Optimisation numérique et contrôle optimal : applications en chimie moléculaire

Adel Ben Haj Yedder

THÈSE

présentée pour l'obtention du titre de

DOCTEUR DE L'ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

Spécialité : Mathématiques et Informatique

par

Adel BEN HAJ YEDDER

Sujet : *Optimisation numérique et Contrôle optimal : applications en chimie moléculaire*

Soutenance le 13 décembre 2002 devant le jury composé de :

Président :	Yvon Maday
Rapporteurs :	Yves Achdou Marc Schoenauer
Examinateurs :	Osman Atabek Jean-Pierre Puel
Directeur de thèse :	Claude Le Bris

À ma famille.

Remerciements

Je voudrais commencer par remercier mon directeur, Claude LE BRIS, qui m'a encadré pendant cette thèse. Il m'a beaucoup apporté par son exigence de clarté et de rigueur ainsi que par son expérience et ses conseils très précieux. Je suis également redevable à Eric CANCES du temps qu'il m'a consacré, de sa pédagogie et des ses encouragements durant cette période.

Le travail que je présente ici n'aurait pu aboutir sans les contributions de Anne AUGER et Claude DION avec qui j'ai eu beaucoup de plaisir à travailler et qui m'ont beaucoup apporté. J'ai eu la chance de travailler avec Osman ATABEK et Marc SCHOENAUER chez qui j'ai particulièrement apprécié leur gentillesse, leur pédagogie et leur disponibilité. Ma collaboration avec Xavier BLANC a été à la fois agréable et enrichissante grâce à son enthousiasme et à son ouverture scientifique. Je remercie aussi Arne KELLER et Stéphane CHELKOWVSKI pour leurs collaborations fructueuses et enrichissantes.

Je tiens à remercier Yvon MADAY d'avoir accepté de présider mon jury de thèse, Yves ACHDOU et Marc SCHOENAUER d'avoir bien voulu rapporter ce travail, Osman ATABEK et Jean-Pierre PUEL d'avoir accepté de faire partie de mon jury.

J'adresse mes sincères remerciements à Bernard LAPEYRE pour son accueil au Cermics et pour ses conseils durant cette période.

Mon séjour au Cermics a été très enrichissant entre autres grâce à la multidisciplinarité de ses équipes et à la convivialité de ses membres. Rares sont les personnes à qui je n'ai pas demandé un jour de l'aide ou posé une question. D'abord un grand merci à Jean-François DELMAS et Benjamin JOURDAIN pour leur conseils en probabilité et à Renaud KERIVEN pour son aide en calcul parallèle. Je remercie aussi infiniment Jacques DANIEL toujours disponible pour répondre aux questions et résoudre les problèmes informatiques. Je ne manquerai pas de remercier Sylvie BERTE, Khadija EL LOUALI et Imane HAMADE pour leur travail formidable au secrétariat et pour leur bonne humeur.

Cette page de remerciements ne serait pas complète si j'oubliais mes colocataires du bureau B412 qui m'ont supporté : Anne pour son aide même pour l'écriture de cette page, François pour son "esprit Ponts" (on se comprend) et Maxime pour ses blagues et sa bonne humeur. Ensuite mes voisins du bureau B411 : Claude pour tout ce qu'il m'a appris en physique, en LaTeX et pour les relectures en anglais (oui

tout ça et même plus !), Linda pour son soutien, Tony qui m'a souvent raccompagné après les "dîners cermics" et Yousra que je remercie particulièrement pour l'organisation de mon pot de thèse. Sans oublier bien sûr les autres thésards et postdocs du quatrième, Adrien, Frédéric, Jennifer, Kengy et Laetitia pour les moments sympathiques partagés ensemble. Je voudrais remercier aussi Maxime BARRAULT, Frédéric LEGOLL, Tony LELIEVRE et Gabriel TURINICI pour les interactions autour de la chimie quantique et de l'optimisation.

Je tiens également à adresser mes remerciements aux autres membres du laboratoire pour les différents échanges que l'on a eu et pour l'ambiance chaleureuse qu'ils ont su créer au Cermics : Geoffray ADDE, Bouhari AROUNA, Gilbert CAPLAIN, Jean-Philippe CHANCELIER, Maureen CLERC, Guy COHEN, Michel COHEN DE LARA, Alexandre ERN, Thérèse GUILBAUD, Olivier JUAN, René LALEMENT, Fabien LE JEUNE, Régis MONNEAU, Nicola MORENI, Jean-François POMMARET, Thierry SALSET et Emmanuel TEMAM.

Et bien sûr une mention particulière va à ma famille et en particulier à mes parents. Je veux leur témoigner toute gratitude et ma reconnaissance pour leur soutien constant.

Enfin, et pour être sûr de n'oublier personne, merci à vous lecteurs pour l'intérêt (scientifique ou autre) que vous portez à mon travail. Un intérêt dont la preuve est que vous avez pris le temps de lire ce document jusqu'à la fin de cette page. Pour ceux qui iront encore plus loin, passons maintenant au vif du sujet,...

Résumé : Ce travail porte, pour l'essentiel, sur l'application des méthodes de contrôle et d'optimisation au contrôle par laser des systèmes moléculaires. La partie principale (Chapitres 1 à 6) est consacrée à l'étude du contrôle par laser de l'orientation moléculaire. Il s'agit de trouver un champ laser capable d'orienter une molécule linéaire le long de l'axe de ce laser. Le premier chapitre présente une introduction générale et passe en revue l'ensemble des méthodes d'optimisation utilisées pour le résoudre. Les chapitres suivants présentent avec plus de détails les différentes méthodes utilisées pour le problème de contrôle par laser (Chapitres 2 et 3) et les principaux résultats obtenus (Chapitres 4, 5 et 6).

Dans le Chapitre 7, on présente des résultats préliminaires sur un autre problème de contrôle par laser utilisant les mêmes outils que ceux présentés dans le premier chapitre. Ce problème concerne l'optimisation de la génération d'harmoniques hautes (HHG) par un atome d'hydrogène excité par un champ laser dans le but de favoriser la création d'un champ laser ultra-court (laser attoseconde).

Dans le Chapitre 8, on présente des outils numériques développés spécifiquement pour traiter des problèmes d'optimisation de géométrie pour la chimie moléculaire. Dans ce problème on cherche à optimiser la position de N particules dont l'énergie d'interaction est donnée (entre autres) par le potentiel de Lennard-Jones.

Enfin, le Chapitre 9 est consacré à des résultats théoriques sur le problème Optimized Effective Potential (OEP) pour la minimisation de l'énergie de Hartree-Fock. Dans ce problème on se pose la question de la validité de la simplification qui consiste à remplacer les équations de Hartree-Fock par des équations aux valeurs propres plus simples.

Abstract : The most important part of this work concerns the application of control and optimization tools to the laser control of molecular systems. The main part of this thesis (Chapters 1 to 6) is devoted to laser control of molecular orientation. Our goal is to find the laser field which orients the molecule along its direction. In the first chapter we present the orientation problem and the different optimization methods we have developed. In the following chapters we give more details about the optimization methods we used (Chapters 2 and 3) and the main results obtained (Chapters 4, 5 and 6).

In Chapter 7 we present another laser control problem using the same optimization tools. In this problem we study the optimization of the High Harmonic Generation (HHG) of an hydrogen atom excited by a laser field. The goal is to create an ultra-short laser field (attosecond laser).

In Chapter 8 we present some numerical tools developed for a geometry optimization problem in molecular chemistry. In this chapter we optimize the position of N particles where the interaction is given (among other cases) by the Lennard-Jones potential.

Finally, in Chapter 9 we give some theoretical results about the Optimized Ef-

fective Potential (OEP) problem for the Hartree-Fock energy minimization. In this problem we ask about the validity of the simplification consisting in replacing the Hartree-Fock equations by some eigenvalues equations of a simplified form.

Sommaire

Introduction générale

Ce travail porte, pour sa grande partie, sur l'application des méthodes de contrôle et d'optimisation au contrôle par laser des systèmes moléculaires. Le contrôle par laser est une branche très active de la physique des lasers et se situe à l'intersection de plusieurs disciplines ; la chimie quantique, la physique quantique et la physique des lasers théorique et expérimentale. Cette branche offre de nouveaux champs de travail, non encore bien explorés, pour les mathématiques appliquées et la simulation numérique.

La partie principale des travaux de cette thèse a porté sur le contrôle par laser de l'orientation moléculaire. Ceci s'inscrit dans le cadre d'un projet lancé en 1999 sur l'*Etude numérique et expérimentale du contrôle des réactions chimiques par laser* financé dans le cadre d'une Action Concertée Incitative Jeunes Chercheurs du Ministère de la Recherche. Ce projet réunit une équipe de chercheurs appartenant à des disciplines très variées : mathématiciens, numériciens, physiciens, chimistes et physiciens expérimentateurs.

Le problème du contrôle par laser de l'orientation moléculaire a été proposé par une équipe de physiciens (O. Atabek, C. M. Dion et A. Keller) du laboratoire de Photo-Physique Moléculaire à Orsay. Cette équipe a mené depuis quelques années des études sur l'alignement et l'orientation des molécules [8–11]. Une collaboration très étroite avec cette équipe a permis de bien poser le problème, de proposer des méthodes numériques et d'analyser les résultats obtenus par la simulation numérique.

La première partie (chapitres 1 à 6) est consacrée à l'étude du contrôle par laser de l'orientation moléculaire. Il s'agit de trouver un champ laser capable d'orienter une molécule linéaire avec l'axe de ce laser. Le premier chapitre présente une introduction générale de ce problème et développe l'ensemble des méthodes d'optimisation utilisées pour le résoudre. Les chapitres 2 à 5 reproduisent des articles publiés respectivement dans *Proceedings of Automatic Differentiation 2000*, *Mathematical Models and Methods in Applied Sciences*, *Physical Review A* (2 articles) et *Proceeding of 41st IEEE Conference on Decision and Control (CDC02)*. Ces chapitres présentent avec plus de détails les différentes méthodes utilisées pour le problème de contrôle par laser (chapitres 2 et 3) et les principaux résultats obtenus (chapitres 4 et 5). Dans le chapitre 2 on présente l'utilisation des outils de différentiation au-

tomatique pour traiter ce problème de contrôle. Dans le chapitre 3 on présente les différentes approches suivies pour traiter ce problème et on détaille les algorithmes déterministes et évolutionnaires utilisés. On présente aussi dans ce chapitre une comparaison des différentes méthodes pour calculer le gradient nécessaire pour les méthodes déterministes. Dans le chapitre 4, on donne le principal résultat du point de vue physique. Ce résultat correspond à un champ laser permettant d'obtenir une orientation selon un mécanisme dit mécanisme de kick. Dans le chapitre 5, on présente la définition de différents critères mesurant l'orientation de la molécule et utilisés par les méthodes d'optimisation. Une étude comparative de ces critères est donnée ensuite en fonction des résultats obtenus pour chaque critère. Le chapitre 6 résume les différents résultats des chapitres précédents et donne en plus des résultats issus d'un modèle de mécanique classique qui approxime le modèle quantique.

La suite de la thèse est consacrée à trois autres sujets. D'abord, au chapitre 7, on présente un autre problème de contrôle par laser utilisant les mêmes outils que ceux présentés dans le premier chapitre. Ce travail est réalisé avec S. Chelkowski du Département de Chimie à Université de Sherbrooke et avec O. Atabek. Ce problème concerne l'optimisation de la génération d'harmoniques hautes (HHG) par un atome d'hydrogène excité par un champ laser. Un des objectifs des générations d'harmoniques hautes est de convertir des lasers standards (UV) en lasers de hautes fréquences (rayon X). Le but de notre étude est de favoriser la création d'un champ laser *attoseconde*, champ très court dont la durée est de l'ordre de quelques $10^{-18}s$. A l'heure actuelle l'objectif (dans cette étude et expérimentalement) est d'atteindre une durée de quelques centaines d'attosecondes.

Dans le chapitre 8, on présente les outils développés pour traiter des problèmes d'optimisation de géométrie dans le cadre d'un travail commun avec X. Blanc au Cermics. Dans ce problème on cherche à optimiser la position de N particules dont l'énergie d'interaction est donnée (entre autres) par le potentiel de Lennard-Jones. En dimension 2, l'énergie de ces particules est donnée par :

$$E(\{X_i\}_{1\leq i\leq N}) = \frac{1}{2}\sum_{i\neq j} W(|X_i - X_j|),$$

où le potentiel de Lennard-Jones W est :

$$W(r) = \frac{1}{r^{12}} - \frac{2}{r^6}.$$

Le traitement numérique est également pour un grand nombre particules (à partir d'une cinquantaine de particules) à cause du grand nombre de minima locaux. D'une part, le problème présente une invariance par translation, par rotation et par permutation des particules. D'autre part, à grande distance le potentiel est quasiment nul ($W(100) \sim 10^{-12}$), ce qui crée d'autres invariances numériques. En effet, si par exemple une particule (ou un groupe de particules) se trouve à une grande distance

des autres particules, l'énergie totale sera inchangée (numériquement) par toute translation qui ne la rapproche pas des autres particules. La méthode développée est une méthode hybride qui combine un algorithme déterministe (gradient conjugué) et un algorithme stochastique (algorithme génétique). Cette méthode utilise également des transformations qui sont issues des résultats théoriques connus sur ce problème.

Enfin, le chapitre 9 est consacré à des résultats théoriques sur le problème Optimized Effective Potential (OEP). Dans ce problème on considère l'énergie de Hartree-Fock donnée par :

$$I^{HF} = inf\{E^{HF}(\phi_1, \ldots, \phi_N), \int_{\mathbb{R}^3} \phi_i \phi_j^* = \delta_{ij} \ , \ 1 \leq i, j \leq N \ , \ \phi_i \in H^1(\mathbb{R}^3, \mathbb{C})\}$$

où

$$\begin{aligned} E^{HF}(\phi_1, \ldots, \phi_N) \ &= \ \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i|^2 - \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{Z}{|x|} |\phi_i|^2 + \frac{1}{2} \int \int_{(\mathbb{R}^3)^2} \frac{\rho(x)\rho(y)}{|x-y|} dxdy \\ &\quad - \frac{1}{2} \int \int_{(\mathbb{R}^3)^2} \frac{|\rho(x,y)|^2}{|x-y|} dxdy, \end{aligned}$$

et

$$\rho(x,y) = \sum_{i=1}^{N} \phi_i(x) \phi_i^*(y) \ \ \rho(x) = \rho(x,x) = \sum_{i=1}^{N} |\phi_i(x)|^2.$$

Les fonctions $(\phi_1, \ldots, \phi_N)$ sont solutions des équations de Hartree-Fock :

$$\Delta \phi_i - \frac{Z}{|x|} \phi_i + (\sum_{j \neq i} |\phi_j|^2 \star \frac{1}{|x|}) \phi_i - (\sum_{j \neq i} \phi_j^* \phi_i \star \frac{1}{|x|}) \phi_j = -\epsilon_i \phi_i.$$

On se pose alors la question de savoir s'il existe un potentiel W (Optimized Effective Potential) pour lequel la minimisation de l'énergie E^{HF} sur les fonctions $(\phi_1, \ldots, \phi_N)$ qui sont solutions des équations :

$$(-\Delta + W)\phi_i = \lambda_i \phi_i, \ i = 1, \ldots, N$$

avec $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ donne la même énergie que le problème initial (I^{HF}). Une étude mathématique de ce problème est menée dans ce dernier chapitre.

Liste des publications parues ou acceptées

[P1] A. Ben Haj Yedder, E. Cancès, and C. Le Bris. Optimal laser control of chemical reactions using automatic differentiation. In George Corliss, Christèle Faure, Andreas Griewank, Laurent Hascoët, and Uwe Naumann (eds.), editors, *Proceedings of Automatic Differentiation 2000 : From Simulation to Optimization*, pages 203–213, New York, 2001. Springer-Verlag.

[P2] C. M. Dion, A. Ben Haj Yedder, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Optimal laser control of orientation : The kicked molecule. *Phys. Rev. A*, 65 :063408, 2002.

[P3] A. Auger, A. Ben Haj Yedder, E. Cancès, C. Le Bris, C. M. Dion, A. Keller, and O. Atabek. Optimal laser control of molecular systems : methodology and results. *Mathematical Models and Methods in Applied Sciences*, 12(9) :1281–1315, 2002.

[P4] A. Ben Haj Yedder, A. Auger, C. M. Dion, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Numerical optimization of laser fields to control molecular orientation. *Phys. Rev. A*, 66 :063401, 2002.

[P5] A. Ben Haj Yedder. Optimal laser control of molecular orientation : some numerical results. In *Proceeding of 41st IEEE Conference on Decision and Control (CDC02)*, Las Vegas, Nevada, USA, 2002 (in press).

Liste des communications dans des conférences

- *Contrôle optimal de l'alignement d'une molécule par un laser (Poster)*, CANUM2000 : 32e Congrès national d'analyse numérique, 5 - 9 juin 2000, Port d'Albret (Landes France).

- *Optimal laser control of chemical reactions using automatic differentiation*, AD 2000 The 3rd International Conference/Workshop on Automatic Differentiation : From Simulation to Optimization, June 19-23, 2000 (Nice, France).

- *Automatic Differentiation for Optimal laser control*, Quantum Control : Mathematical and Numerical Challenges, 6-11 octobre 2002 (Montréal, Québec, Canada).

- *Optimal laser control of molecular orientation : some numerical results*, IEEE 2002 Conference on Decision and Control, December 10 - 13, 2002 (Las Vegas, Nevada, USA).

Chapitre 1

Introduction au contrôle par laser : contexte et méthodologie

1.1 Présentation

L'idée de contrôler les réactions chimiques par des lasers est née il y a plus de 40 ans. Cependant, elle n'a été mise réellement en pratique que depuis une dizaine d'années avec la naissance des nouvelles générations de laser. En effet, ce n'est qu'après l'apparition des sources laser femtosecondes de haute intensité pouvant générer des impulsions très courtes (1 femtoseconde = $10^{-15}s$) que les chercheurs ont pu s'intéresser non seulement à l'étude de l'interaction laser-matière, mais aussi à la possibilité de contrôler les atomes et les molécules à l'aide d'un champ laser. Ainsi, l'utilisation des lasers en vue du contrôle des molécules est devenue aujourd'hui une réalité expérimentale [1], donnant accès à la compréhension des mécanismes des réactions chimiques ou à la synthèse de produits impossibles à obtenir par les techniques habituelles par exemple.

Rappelons ici quelques chiffres, afin d'avoir une idée des ordres de grandeur dans le monde de la physique quantique et des lasers femtosecondes. L'échelle en espace est celle d'une molécule, elle est de l'ordre de quelques angströms ($10^{-10}m$), et l'échelle de temps est celle des vibrations des liaisons moléculaires, de l'ordre de la femtoseconde ($10^{-15}s$). A l'heure actuelle les simulations numériques des équations de la physique quantique (basées sur l'équation de Schrödinger ou sur des modèles l'approchant) se font sur des durées de l'ordre de la picoseconde ($10^{-12}s$). Du point de vue du contrôle, ceci implique que le contrôle ne peut se faire qu'en boucle ouverte. En effet le matériel expérimental ne peut pas réagir aussi rapidement. En revanche, ces échelles de temps très courtes ont donné l'idée à H. Rabitz [18] de faire du *contrôle expérimental*. Au lieu de simuler un système moléculaire (ce qui demande beaucoup de temps de calcul) on laisse le système réagir (ce qui est quasiment instantané !) et on mesure le résultat. On peut ainsi réaliser des milliers d'expériences par minute

et les mesures obtenues peuvent servir dans une boucle d'optimisation.

Parmi les exemples de problèmes de contrôle par laser on peut donner l'exemple de la rupture sélective d'une liaison chimique. Le schéma typique est le suivant :

$$ABC \longrightarrow \left\{ \begin{array}{ccc} A & + & BC \\ AB & + & C \end{array} \right.$$

où l'on veut favoriser la réalisation de la première dissociation par rapport à la seconde par exemple. Le but est de trouver un champ laser optimisé qui a pour effet de briser la liaison $A - B$ alors qu'un champ laser non optimisé a tendance à briser l'autre liaison ou les deux liaisons simultanément. Ce schéma peut également modéliser une réaction chimique qui produit à la fois $A + BC$ d'une part et $AB + C$ d'autre part. Le but, dans ce cas, est de trouver le champ laser qui permet d'améliorer le rendement de cette réaction en favorisant la production de $A + BC$ plutôt que celle de $AB + C$.

Le choix du problème du contrôle par laser de l'orientation moléculaire se justifie par plusieurs arguments. Premièrement, cette étude permet de mieux comprendre les mécanismes d'interaction laser-molécule et d'aider dans la découverte de nouveaux mécanismes. Deuxièmement, le contrôle de l'orientation est une étape vers le contrôle des réactions chimiques. En effet, orienter les molécules avant une réaction revient à les préparer de façon à ce que la réaction se réalise spontanément ensuite. Troisièmement, des résultats récents ont montré la faisabilité expérimentale du contrôle de l'alignement (l'orientation est une étape après l'alignement) par un champ laser et le rôle de l'alignement pour le contrôle des réactions [19, 20, 26].

Présentons maintenant un peu plus le modèle du système étudié et le problème de contrôle avant de le détailler dans la section suivante 1.2. Le but de l'étude est de trouver un champ laser capable d'orienter une molécule linéaire avec son axe de polarisabilité. Le système moléculaire soumis au champ laser $\vec{\mathcal{E}}$ est gouverné par l'équation de Schrödinger dépendante du temps :

$$i\hbar \frac{\partial \psi}{\partial t} = H_0 \, \psi + \vec{\mathcal{E}}(t) \cdot \vec{D}\left(\vec{\mathcal{E}}(t)\right) \, \psi, \tag{1.1}$$

complétée de la condition initiale $\psi(t = 0) = \psi_0$. Dans cette équation on suppose que la fonction d'onde ψ, qui représente l'état de la molécule, ne dépend que des coordonnées des atomes composant la molécule. La présence des électrons est prise en compte par un potentiel effectif qui agit sur les atomes et qui est contenu dans l'hamiltonien H_0 du système libre (sans la présence du champ laser). L'opérateur $\vec{D}\left(\vec{\mathcal{E}}(t)\right)$ représente le moment dipolaire de la molécule en présence d'un champ électrique externe $\vec{\mathcal{E}}(t)$. Dans une approximation au premier ordre, ce moment s'écrit sous la forme

$$\vec{D}\left(\vec{\mathcal{E}}(t)\right) = \vec{\mu}_0 + \overline{\overline{\alpha}}\vec{\mathcal{E}},$$

où $\vec{\mu}_0$ désigne le moment dipolaire permanent de la molécule et où $\overline{\overline{\alpha}}\vec{\mathcal{E}}$ représente le moment dipolaire induit par le champ $\vec{\mathcal{E}}$ ($\overline{\overline{\alpha}}$ est le tenseur de polarisabilité). Des modèles plus sophistiqués utilisent un développement à un ordre supérieur pour $\vec{D}(\vec{\mathcal{E}}(t))$ ou encore une dépendance complète de la fonction d'onde ψ et de l'hamiltonien H par rapport à toutes les coordonnées des atomes et des électrons du système moléculaire. Ces modèles sont encore hors de portée des simulations numériques actuelles.

Pour poser le problème de contrôle, il est nécessaire de définir une fonctionelle de coût dont la minimisation traduit l'objectif physique à atteindre (orienter la molécule). Par exemple si l'on considère dans un cas simple que la fonction d'onde ψ, solution de l'équation (1.1), dépend uniquement du temps t et de l'angle θ, angle entre l'axe de la molécule et la direction du champ laser, on définit la fonctionelle de coût $J(\mathcal{E})$ de la façon suivante :

$$J(\mathcal{E}) = \frac{1}{T}\int_{t=0}^{t=T}\int_{\theta=0}^{\theta=\pi} |\psi(t,\theta)|^2 \cos\theta \, \sin\theta \, d\theta \, dt. \tag{1.2}$$

Le problème de contrôle ainsi posé diffère des problèmes de contrôle rencontrés usuellement dans d'autres domaines de la physique. D'abord, il s'agit d'un problème de *contrôle optimal* et non de *contrôle exact*. Il ne s'agit pas d'amener le système d'un état initial à un état cible Ψ_T à l'instant final T. En effet, on ne connaît pas un état Ψ_T qui réalise l'orientation et qui plus est , on n'est même pas sûr de l'existence d'un tel état. Une autre particularité de ce problème est que le contrôle $\vec{\mathcal{E}}$ multiplie l'état ψ, on parle de *contrôle bilinéaire*. Ceci a pour effet de rendre le problème très difficile sur le plan théorique. En effet, les résultats théoriques sur le contrôle bilinéaire en dimension infinie sont très rares depuis les travaux de Ball et Slemrod [4]. Dans le cas de la dimension finie, il existe quelques résultats théoriques sur la contrôlabilité exacte des systèmes approximant l'équation (1.1). On cite par exemple les travaux de G. Turinici et H. Rabitz [29–31]. On possède encore moins de résultats théoriques sur le problème de contrôle optimal : par exemple dans [6], on montre l'existence d'un contrôle optimal sur un cas simplifié de l'équation (1.1). Dans cette thèse on s'est principalement concentré sur les aspects numériques du problème de contrôle.

On note enfin que le contrôle $\vec{\mathcal{E}}$ est *distribué en temps* en non en espace. Ce point n'induit aucune perte de généralité, le cas où le champ est distribué à la fois en espace et en temps ($\vec{\mathcal{E}}(t,x)$) se traite de manière semblable au prix de calculs plus compliqués. D'autre part, cette approximation est raisonnable pour un système moléculaire de petite taille où la variation du champ $\vec{\mathcal{E}}$ est négligeable à l'échelle du système.

1.2 Le problème physique

1.2.1 Le système moléculaire

Le système étudié est celui d'une molécule linéaire soumise à un champ laser intense. Deux molécules sont considérées : la molécule HCN (Cyanure d'Hydrogène) et la molécule LiF (Fluorure de Lithium). Dans leur état fondamental, ces deux molécules sont linéaires tant que la fréquence du laser reste hors résonance avec la fréquence de flexion de la molécule, ce qui est le cas pour les lasers utilisés. L'étude physique détaillée de ce système est donnée par DION dans [8]. Dans la suite, on ne présente que les éléments nécessaires à la définition du problème de contrôle.

On repère la molécule à l'aide des coordonnées de Jacobi ($\mathbf{R} = (R, r), \theta, \varphi$) (voir Figure 1.1). Dans l'approximation de Born-Oppenheimer, l'hamiltonien H_0 du système libre s'écrit

$$H_0 = H_{vib}(\mathbf{R}) + H_{rot}(\mathbf{R}, \theta, \varphi) + V(\mathbf{R}),$$

où $H_{vib} + H_{rot}$ représente l'opérateur cinétique avec

$$H_{vib}(\mathbf{R}) = -\frac{\hbar^2}{2\mu_{HCN}} \frac{1}{R^2} \frac{\partial}{\partial R} \left(R^2 \frac{\partial}{\partial R} \right) - \frac{\hbar^2}{2\mu_{CN}} \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right)$$

et

$$H_{rot}(\mathbf{R}, \theta, \varphi) = B \left[\frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2} \right],$$

où la constante de rotation B est donnée par

$$B = -\frac{\hbar^2}{2(\mu_{HCN} R^2 + \mu_{CN} r^2)}$$

et où $V(\mathbf{R})$ représente le potentiel issu de l'interaction entre les noyaux et les électrons (dans leur état fondamental). Dans ces formules, μ_{CN} et μ_{HCN} sont les masses réduites données par :

$$\mu_{CN} = \frac{m_C m_N}{m_C + m_N} \quad \text{et} \quad \mu_{HCN} = \frac{m_H (m_C + m_N)}{m_H + m_C + m_N}.$$

Le moment dipolaire de la molécule en présence du champ laser s'écrit en utilisant une approximation du premier ordre :

$$D\left(\mathcal{E}(t)\right) = -\mu_0(R, r)\cos\theta - \frac{\mathcal{E}(t)}{2} \left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta \right]$$

où μ_0 désigne le moment dipolaire permanent. Les cœfficients $\alpha_{\parallel}$ et $\alpha_{\perp}$ sont respectivement les composantes parallèle et perpendiculaire du tenseur de polarisibilité $\overline{\overline{\alpha}}$.

Lorsque l'axe (Oz) est choisi parallèle à l'axe moléculaire, le tenseur de polarisibilité $\overline{\overline{\alpha}}$ est diagonal et est donné par $\alpha_{zz} = \alpha_{\parallel}$ et $\alpha_{xx} = \alpha_{yy} = \alpha_{\perp}$.

La forme générale de l'hamiltonien du système en présence du champ laser est donc :

$$H(\mathbf{R}, \theta, \varphi, t) = H_{vib}(\mathbf{R}) + H_{rot}(\mathbf{R}, \theta, \varphi) + V(\mathbf{R}) + H_{laser}(\mathbf{R}, \theta, \varphi, t) \tag{1.3}$$

où l'hamiltonien $H_{laser}(\mathbf{R}, \theta, \varphi, t)$ décrivant l'interaction de la molécule avec le champ laser est donné par :

$$\begin{aligned} H_{laser}(\mathbf{R}, \theta, \varphi, t) &= \mathcal{E}(t) \cdot D(\mathcal{E}(t)) \\ &= -\mu_0(R, r)\mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta\right]. \end{aligned}$$

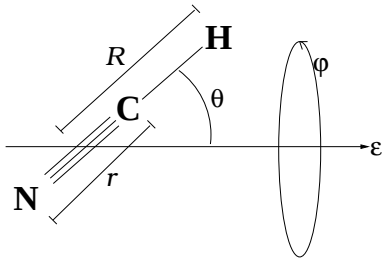


FIG. 1.1 – Le modèle de la molécule HCN.

La simulation numérique du système complet est très coûteuse en temps de calcul, ce qui rend difficile le traitement du problème de contrôle du système dans toute sa généralité. Par conséquent, on s'intéresse au modèle simplifié du *rotateur rigide* caractérisé par une dépendance du problème par rapport aux seules variables angulaires θ et ϕ. La symétrie du problème par rapport à l'axe de polarisation du champ laser permet de séparer la variable ϕ du problème et de se ramener qu'à une seule la dépendance en θ. Le hamiltonien (1.3) devient alors :

$$H = H(\theta, t) = H_{rot}(\theta) + H_{laser}(\theta, t) \tag{1.4}$$

avec

$$H_{rot}(\theta) = B\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)\right]$$

et

$$H_{laser}(\theta, t) = -\mu_0(R, r)\mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta\right]$$

TAB. 1.1 – Paramètres des molécules HCN et LiF.

molécule	B (a.u.)	μ_0 (a.u.)	$\alpha_\parallel$ (a.u.)	$\alpha_\perp$ (a.u.)	T_{rot} (ps)
HCN	6.6376×10^{-6}	1.1413	20.055	8.638	11.45
LiF	5.9173×10^{-6}	2.5933	9.061	9.218	12.84

où les variables R et r sont fixées à leur valeur d'équilibre. Le Tableau 1.1 résume les paramètres des molécules HCN et LiF dans leur état fondamental.

L'équation de Schrödinger 1.1 dépendant uniquement de la variable θ ainsi obtenue est la suivante :

$$\begin{cases} i\hbar\dfrac{\partial\psi}{\partial t}(\theta,t) = H(\theta,t)\,\psi, \\ \psi(t=0) = \psi_0. \end{cases} \tag{1.5}$$

A l'instant $t = 0$, et dans le cas général, la molécule est prise dans son état fondamental. L'état initial ψ_0 est donc donné par la première harmonique sphérique $Y_{0,0}$. Dans la Section 1.2.4, on présente un modèle plus proche de la réalité expérimentale correspondant à des conditions initiales différentes. Cette équation est résolue numériquement par deux programmes en Fortran écrit par DION [8] utilisant deux approches différentes. La première approche utilise la méthode de décomposition d'opérateurs (operator splitting) [14] couplée avec une transformée de Fourier rapide (FFT) pour la partie cinétique comme expliquée dans [7, 24]. La seconde méthode consiste à développer la fonction d'onde $\psi(\theta,\varphi,t)$ sur une base d'harmoniques sphériques

$$\psi(\theta,\varphi,t) = \sum_{J=0}^{J_{max}} \sum_{M=-J}^{J} c_{J,M}(t) Y_{J,M}(\theta,\varphi),$$

où les fonctions $Y_{J,M}(\theta,\varphi)$ sont les fonctions propres de l'opérateur H_{rot}. Ainsi, on se ramène à la résolution d'un système d'équations ordinaires couplées sur les cœfficients $c_{J,M}$ qui est résolu par un schéma de Runge-Kutta d'ordre 4.

La mesure de l'orientation de la molécule à un instant t est donnée par un critère instantané $j(t)$ (voir détail dans [15]) défini par :

$$j(t) = \langle \cos\theta \rangle = \int_0^\pi \cos\theta \;\; \mathcal{P}(\theta,t) \sin\theta \, d\theta, \tag{1.6}$$

où $\mathcal{P}(\theta,t)$ représente la distribution angulaire de la molécule. Dans le modèle du rotateur rigide, l'expression de cette distribution se réduit à $\mathcal{P}(\theta,t) = \|\psi\|^2_{\mathbb{C}}$ avec $\|\psi\|^2_{\mathbb{C}}$ le carré de la norme de la fonction (complexe) ψ solution du système (1.5). Le critère instantané devient alors

$$j(t) = \int_0^\pi \cos\theta \;\; \|\psi\|^2_{\mathbb{C}} \sin\theta \, d\theta. \tag{1.7}$$

Le critère instantané $j(t)$ prend ses valeurs dans l'intervalle $[-1, 1]$, les valeurs -1 et 1 correspondent respectivement à une molécule orientée dans le même sens que le champ laser et à une molécule orientée dans le sens opposé. Notons ici que le critère instantané $j(t)$ dépend des paramètres du champ laser $\mathcal{E}$ par l'intermédiaire de la fonction d'onde ψ.

1.2.2 Choix du champ laser

Le champ laser recherché est considéré comme la superposition de N champs élémentaires (N pouvant aller jusqu'à 10) polarisés suivant le même axe (voir Figure 1.2). Bien que sur le plan expérimental N ne puisse dépasser 3, on a choisi de s'autoriser un plus grand nombre de lasers afin d'élargir l'espace de recherche et ainsi trouver éventuellement des solutions physiquement intéressantes mais non réalisables dans un futur proche. La superposition d'un grand nombre de lasers présente deux inconvénients majeurs. Premièrement, ellle rend la recherche d'une solution plus difficile car l'espace de recherche devient beaucoup plus grand. Deuxièmement, comme on le verra par la suite, les solutions obtenues avec 10 lasers sont plus difficiles à comprendre et à interpréter physiquement. Ainsi, dans beaucoup de simulations on a considéré la superposition de 2 ou 3 champs lasers élémentaires. De telles simulations ont permis de trouver des solutions physiquement intéressantes comme notamment le *champ kick* présenté dans [P2].

Le champ laser s'écrit sous la forme :

$$\mathcal{E}(t) = \sum_{n=1}^{N} \mathcal{E}_n(t) \sin\left(\omega_n t + \phi_n\right).$$

L'enveloppe $\mathcal{E}_n(t)$, qui obéit à des contraintes expérimentales, est modélisée par :

$$\mathcal{E}_n(t) = \begin{cases} 0 & \text{si} \quad t \leq t_{0n}, \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t-t_{0n}}{t_{1n}-t_{0n}}\right)\right] & \text{si} \quad t_{0n} \leq t \leq t_{1n}, \\ \mathcal{E}_{0n} & \text{si} \quad t_{1n} \leq t \leq t_{2n}, \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t_{3n}-t}{t_{3n}-t_{2n}}\right)\right] & \text{si} \quad t_{2n} \leq t \leq t_{3n}, \\ 0 & \text{si} \quad t \geq t_{3n}. \end{cases} \tag{1.8}$$

Chaque champ est caractérisé par 7 paramètres : sa fréquence ω_n, sa phase relative ϕ_n, son amplitude $\mathcal{E}_{0n}$ et les quatre instants déterminant la forme de son enveloppe (l'origine t_{0n}, le temps d'allumage $t_{1n} - t_{0n}$, le plateau $t_{2n} - t_{1n}$, et le temps d'extinction $t_{3n} - t_{2n}$). Ces paramètres doivent satisfaire les contraintes suivantes :

$$\begin{aligned} &\mathcal{I} = \epsilon_0 c \mathcal{E}^2/2 \leq 3 \times 10^{13}\ \mathrm{W/cm^2}, \\ &500 \leq \omega_n \leq 4000\ \mathrm{cm^{-1}}, \\ &t_{0n} \leq t_{1n} \leq t_{2n} \leq t_{3n} \leq 1.7\ \mathrm{ps}. \end{aligned}$$

Les temps d'allumage et d'extinction doivent aussi satisfaire une contrainte technique qui impose que ces temps soient plus longs d'au moins 5 fois la période du laser (le passage d'un laser éteint à un laser d'intensité $\mathcal{E}_{0i}$ ne peut se faire instantanément).

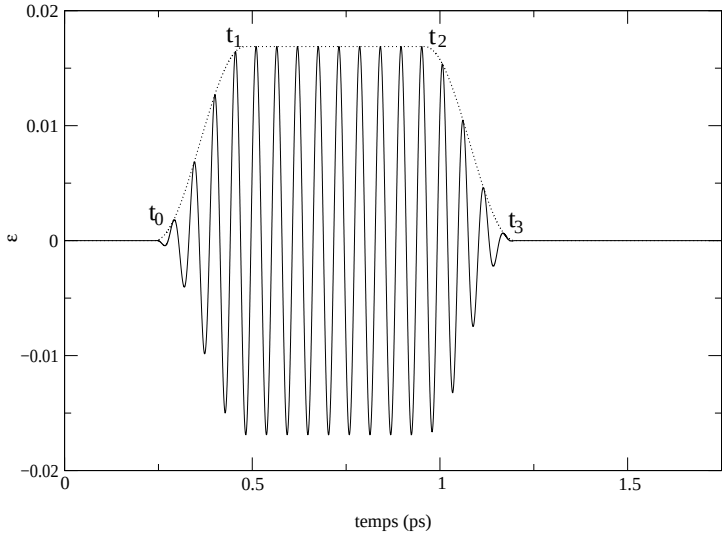
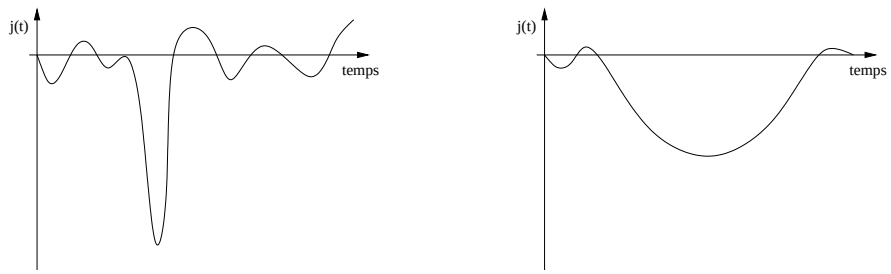


FIG. 1.2 – Un champ laser élémentaire $\mathcal{E}_i(t)$.

1.2.3 Choix du critère à optimiser

Du point de vue physique, obtenir une bonne orientation peut se traduire de plusieurs manières différentes. D'une part, on peut s'intéresser à la valeur absolue du critère instantané $|\langle\cos\theta\rangle(t)|$ (défini par l'équation 1.7) que l'on veut rendre proche de 1 à un instant quelconque. D'autre part, on peut chercher à obtenir une orientation d'une durée la plus longue possible. La longueur de la durée de l'orientation est à comparer à la période rotationnelle de la molécule (voir Tableau 1.1). La Figure 1.3 illustre l'allure typique du critère instantané dans ces deux cas. Enfin, les physiciens s'intéressent aussi bien à l'orientation en présence du champ laser qu'à l'orientation après l'extinction du dernier laser. La formulation du critère à optimiser a évolué au cours du temps en fonction des différents résultats obtenus et après des interactions avec les physiciens. Ainsi, plusieurs critères ont été testés. On peut les classer en deux catégories : les critères simples, qui ne prennent en compte que l'un des deux objectifs, et les critères hybrides dont le but est de trouver un compromis entre l'efficacité de l'orientation et sa durée. Dans la suite on présente un ensemble de critères testés donnant chacun un résultat spécifique que l'on présente dans la Section 1.4. On trouve dans [P4] une présentation de certains de ces critères et des



(a) (b)

FIG. 1.3 – Schéma d'un critère $j(t)$ donnant une bonne orientation de courte durée (a) et d'un critère $j(t)$ donnant une moins bonne orientation mais pour une durée plus longue (b).

résultats obtenus.

Afin de rendre les formules plus compactes on adopte les notations suivantes : T_f désigne l'instant final correspondant à l'extinction du dernier laser, T_{rot} désigne la période rotationnelle de la molécule et $[T_a, T_b]$ désigne l'intervalle de temps pendant lequel on veut réaliser l'orientation. Cet intervalle pourra désigner l'intervalle $[0, T_f]$, l'intervalle $[0, T_f + T_{rot}]$ ou le plus fréquent, l'intervalle $[T_f, T_f + T_{rot}]$. Tous les critères sont formulés pour que leur optimisation revienne à leur minimisation. Dans la suite ces critères seront notés $J(\mathcal{E})$ où $\mathcal{E}$ représente les paramètres du champ laser.

Critères simples :

- $J_1 = j(T_{fin})$: on veut obtenir une bonne orientation à la fin de l'impulsion laser.
- $J_2 = \min_{t \in [T_a, T_b]} j(t)$: on cherche à optimiser l'orientation à un instant quelconque dans l'intervalle $[T_a, T_b]$.
- $J_3 = -\dfrac{\tau}{T_{rot}}$ où τ est la durée pendant laquelle l'orientation est restée suffisamment bonne. Plus précisément, τ est la longueur de l'intervalle connexe $I \subset [T_a, T_b]$ telle que $\forall t \in I \quad \dfrac{J_1}{\sqrt{2}} \leq j(t) \leq J_1$ (voir Figure 1.4). Un tel critère, pris seul ne permet pas d'obtenir une bonne orientation car J_1 peut être faible.

Critères hybrides :

- $J_4 = J_2 + J_3 + |J_2 - J_3|$: ce critère tend à minimiser à la fois J_2 (donc la mesure de l'orientation) et J_3 (donc maximiser sa durée). Le terme $|J_2 - J_3|$ est un terme de pénalisation pour assurer que J_2 et $-J_3$ sont simultanément minimisés.
- $J_5 = \dfrac{1}{T_b - T_a} \displaystyle\int_{T_a}^{T_b} j(t)dt$: ce critère qui représente une "moyenne" de l'orientation sur l'intervalle $[T_a, T_b]$ se révèle sans intérêt dans certains cas. En effet,

lorsque $T_b - T_a = T_{rot}$ et $T_a \geq T_f$, ce critère vaut toujours zéro à cause de la symétrie et la périodicité de $j(t)$. Ceci nous a amené à définir les critères J_6 et J_7 suivants.

- $J_6 = -\dfrac{1}{T_b - T_a}\displaystyle\int_{T_a}^{T_b} j^2(t)dt$: cette "moyenne" de $j^2(t)$ n'exclut pas les fortes oscillations du critère instantané.
- $J_7 = -\dfrac{1}{T_b - T_a}\displaystyle\int_{T_a}^{T_b} \mathcal{C}^2(t)dt$: avec $\mathcal{C}(t) = \begin{cases} 0.1j(t) & \text{si} \quad |j(t)| < 0.4, \\ j(t) & \text{sinon.} \end{cases}$
  Cette définition donne une importance plus grande aux intervalles de temps pendant lesquels $j(t)$ est inférieur à une certaine valeur (ici 0.4 par exemple) en réduisant sa valeur sur les autres intervalles (multiplication par un facteur 0.1 dans cet exemple). Ceci a pour conséquence de favoriser les champs laser donnant un critère instantanés $j(t)$ présentant peu d'oscillations.

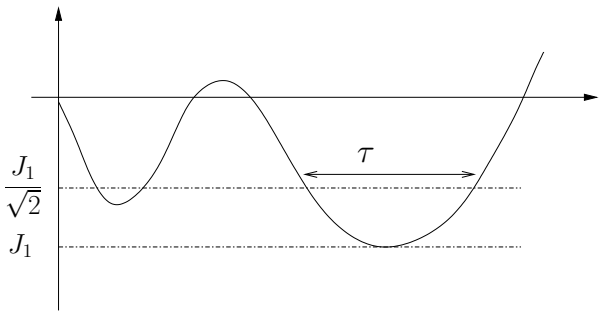


FIG. 1.4 – Construction du critère hybride J_3.

1.2.4 Effet de la température

L'étude du système moléculaire présentée jusqu'ici repose sur un modèle idéal où la fonction d'onde est solution du système composé de l'équation (1.5) et de la condition initiale $\psi_0 = Y_{0,0}$. Ce modèle ne traduit pas complètement la réalité expérimentale. En effet, il est adapté au traitement d'une molécule isolée ou d'un ensemble de molécules à la température de $0K$ ce qui est impossible à réaliser expérimentalement. Afin de s'approcher des conditions expérimentales on doit considérer une distribution d'états initiaux excités, à une température $T > 0$, répartis selon la fonction de répartition

$$\mathcal{Q}(J) = (2J+1)\exp\left[-\frac{BJ(J+1)}{k_B T}\right],$$

où k_B désigne la constante de Boltzmann et T désigne la température. Ceci nous mène à résoudre $N_J = 2J_{max}(J_{max}+2)$ systèmes donnés par l'équation (1.5) et par

la condition initiale $\psi_0 = Y_{J,M}$. La mesure de l'orientation est alors donnée par le critère instantané moyen

$$\langle j \rangle(t) = \langle\langle \cos\theta \rangle\rangle(t) = \mathcal{Q}^{-1} \sum_{J}^{J_{max}} (2J+1) \exp\left[-\frac{BJ(J+1)}{k_B T}\right] \sum_{M=-J}^{J} \langle \cos\theta \rangle_{J,M}(t),$$

où

$$\mathcal{Q} = \sum_{J}^{J_{max}} \mathcal{Q}(J) = \sum_{J}^{J_{max}} (2J+1) \exp\left[-\frac{BJ(J+1)}{k_B T}\right]$$

et où $\langle \cos\theta \rangle_{J,M}(t)$ dénote le critère instantané pour la solution du système avec l'état initial $\psi_0 = Y_{J,M}$. Plusieurs études [21, 22, 25] ont montré que le degré d'alignement ou d'orientation se dégrade quand la température augmente comme le montre la Figure 1.5. Ainsi, un champ laser optimisé pour une température $T = 0K$ (c'est à dire avec un seul état initial $\psi_0 = Y_{0,0}$) ne peut pas donner une bonne orientation pour une distribution à une température $T > 0$. Une première approche a consisté à trouver un champ optimisé pour l'état initial dominant dans la distribution à $T > 0K$ (dans le cas d'une température $T = 5K$, cet état est $\psi_0 = Y_{1,0}$) et à utiliser le champ trouvé pour toute la distribution. Cette approche n'est pas efficace car le critère instantané moyen $\langle\langle \cos\theta \rangle\rangle$ se dégrade également et l'orientation se perd sous l'effet de la moyenne. Dans une deuxième approche, plus fructueuse, on a utilisé directement le critère instantané moyen $\langle\langle \cos\theta \rangle\rangle$ dans la définition du critère à optimiser. Par exemple, le critère J_2 devient

$$\langle J_2 \rangle = \min_{t \in [T_a, T_b]} \langle j \rangle(t).$$

Cette approche revient à trouver le champ optimal qui oriente d'une façon synchrone les différentes molécules, excitées dans leur état initial selon la fonction de répartition $\mathcal{Q}(J)$. Par le terme synchrone, on exprime le fait qu'il existe un instant t pour lequel la "majorité" des molécules sont bien orientées.

1.3 Méthodologie

Dans cette partie, on présente les différentes méthodes d'optimisation utilisées pour le problème d'orientation moléculaire présenté plus haut. Ces méthodes ont pour objectif de trouver la ou les solutions $\mathcal{E}$, représentant les paramètres du champ laser, qui minimise un critère $J(\mathcal{E})$. La Figure 1.6 illustre l'approche pour trouver le champ laser optimal réalisant l'orientation de la molécule. On distingue principalement deux familles d'algorithmes : les algorithmes déterministes du type gradient et les algorithmes évolutionnaires. Les algorithmes déterministes sont des méthodes locales très rapides à converger mais présentent l'inconvénient de rester bloqués dans

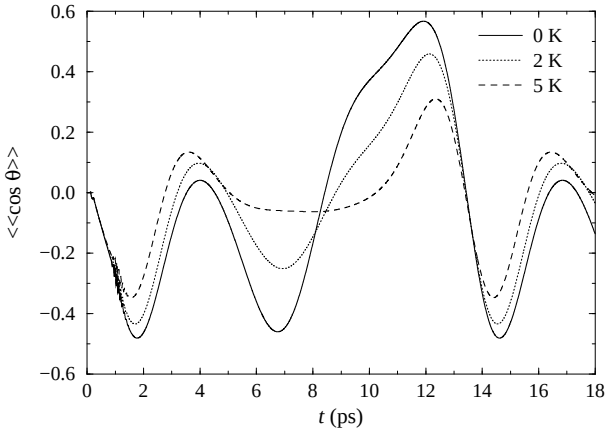


FIG. 1.5 – Effet de la température : le critère instantané moyen $\langle\langle \cos\theta \rangle\rangle$ pour la molécule LiF soumise à un champ laser optimisé pour $T = 0K$.

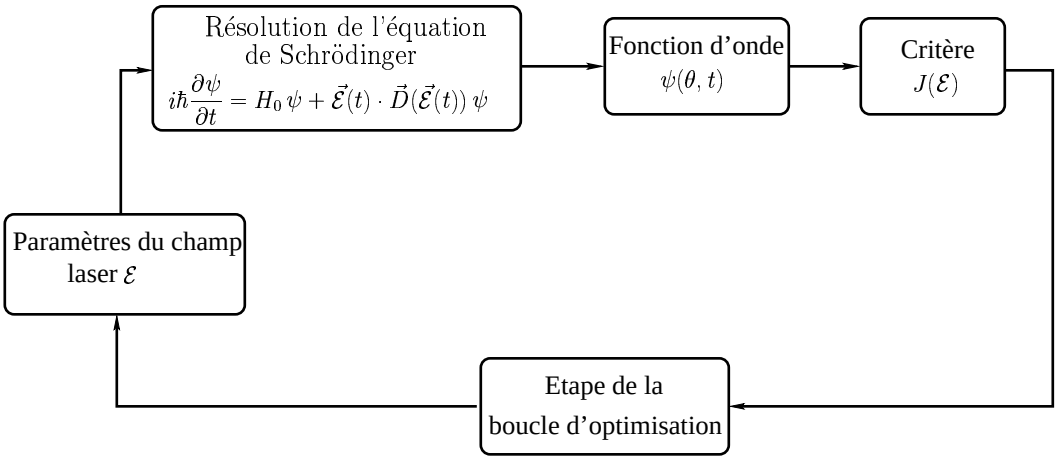


FIG. 1.6 – Schéma pour trouver le champ laser optimal.

des minima locaux. Les algorithmes évolutionnaires sont des méthodes globales de type stochastique et sont beaucoup moins sensibles aux minima locaux. Ces algorithmes sont dit d'ordre zéro car ils ne nécessitent que l'évaluation de la fonction à minimiser. Ce point les rend moins exigeants et plus souples lorsque l'on souhaite modifier la fonction à minimiser par exemple. Les algorithmes évolutionnaires requièrent un grand nombre d'évaluation de fonction ce qui peut les ralentir considérablement surtout quand la fonction à optimiser est coûteuse en temps de calcul.

Afin de tirer profit de l'efficacité des méthodes déterministes à trouver les minima locaux et de la capacité des algorithmes évolutionnaires à effectuer une recherche globale, on a testé des méthodes hybrides combinant les deux approches. Une de ces méthodes consiste à d'utiliser un algorithme génétique avec une mutation effectuée par un algorithme de gradient conjugué (voir Section 1.3.2.6). Cet algorithme n'est pas efficace sur le problème de l'orientation. En revanche, il a permis d'obtenir des résultats intéressants sur le problème de l'optimisation du potentiel Lennard-Jones présenté dans le chapitre 8.

1.3.1 Algorithmes déterministes

Dans cette partie, on présente quelques méthodes déterministes, de type gradient, utilisées pour résoudre le problème de l'orientation. Ces méthodes supposent que la fonction à minimiser est différentiable ce qui restreint la liste des critères présentés dans la Section 1.2.3 et exclut par exemple un critère de type J_2. Les algorithmes testés sont l'algorithme de gradient conjugué non linéaire de Polak-Ribière avec recherche linéaire de Wolfe ou de Goldstein-Price (GCNL) et l'algorithme BFGS avec recherche linéaire du même type. Ces algorithmes sont présentés dans la Section 1.3.1.1. Dans la Section 1.3.1.2 on présente différentes méthodes pour calculer le gradient de la fonction à minimiser nécessaire pour ces méthodes. La Section 1.3.1.3 donne ensuite un exemple de calcul de gradient utilisant l'outil de différentiation automatique O∂yssée .

1.3.1.1 Les algorithmes utilisés

On présente ici le principe des méthodes déterministes de type gradient. Pour une présentation détaillée de ces méthodes en particulier et des méthodes déterministes en général on renvoie à [5].

Pour minimiser une fonction $f(x)$ avec $x \in \mathbb{R}^n$, les algorithmes d'optimisation de type gradient cherchent un point $\overline{x}$ tel que $\nabla f(\overline{x}) = 0$ en construisant une suite minimisante $\{x_k\}$ tel que $liminf \nabla f(x_k) = 0$.

Le schéma général de ces algorithmes est le suivant :

- `Etape 0` choisir un point initial x_0,
- `Etape 1` test d'arrêt : $|\nabla f(x_k)| \leq \epsilon$?,
- `Etape 2` calcul d'une direction de descente en cherchant une direction d_k minimisant un problème approché du type

$$(P_k) : \min_{\{d, \|d\| \leq a\}} f(x_k) + (g(x_k), d) \left[+ \frac{1}{2}(M_k^{-1} d, d) \right],$$

- `Etape 3` calcul d'un pas (Recherche linéaire) : trouver un pas t_k qui fait décroître f suffisamment en cherchant à minimiser $q(t) = f(x_k + t d_k)$,
- `Etape 4` faire $x_{k+1} = x_k + t_k d_k$ et $k = k + 1$ puis revenir à l'`Etape 1`.

Calcul de la direction de descente :
La première idée pour calculer la direction de descente est de prendre la direction donnée par plus forte pente : $d_k = -\nabla f(x_k)$. Cette idée est à éviter car elle est très lente mais elle à la base des autres méthodes.

L'algorithme de *gradient conjugué linéaire* construit les directions de descente de la façon suivante :

$$\begin{cases} d_1 = -g_1, \\ d_{k+1} = -g_{k+1} + c_k d_k \text{ avec } c_k = \dfrac{|g_{k+1}|^2}{|g_k|^2}. \end{cases}$$

Dans le cas d'une fonction quadratique cet algorithme converge en au plus n itérations pour le pas $t_k = -\dfrac{(g_k, d_k)}{(Ad_k, d_k)}$. L'extension de cet algorithme au cas non linéaire permet la construction des directions de descente comme suit :

$$\begin{cases} d_1 = -g_1, \\ d_k = -g_k + c_{k-1} d_{k-1} \text{ et si } (d_k, g_k) \geq 0 \text{ prendre } d_k = g_k. \end{cases}$$

Le cœfficient c_{k-1} est donné par $c_{k-1} = \dfrac{|g_k|^2}{|g_{k-1}|^2}$ pour l'algorithme de Fletcher-Reeves (même choix que pour le cas quadratique) et par $c_{k-1} = \dfrac{(g_k - g_{k-1}, g_k)}{|g_k|^2}$ pour l'algorithme de Polak-Ribière. En pratique c'est ce dernier algorithme qui est utilisé.

L'algorithme de *Newton* utilise le hessien pour la construction de la direction de descente : $d_k = -M_k \nabla f(x_k)$ avec $M_k = (\nabla^2)^{-1} f(x_k)$. Quand on ne dispose pas du hessien on peut utiliser les algorithmes de *Quasi-Newton* dont le principe consiste à approcher l'inverse du hessien en construisant une suite de matrices M^k symétriques, de préférence définie positives et vérifiant $y_k = M^k s_k$ avec $y_k = g_{k+1} - g_k$ et $s_k = x_{k+1} - x_k$. L'exemple le plus utilisé des algorithmes de Quasi-Newton est

l'algorithme *BFGS* qui construit la matrice M_k de la manière suivante :

$$\begin{cases} M_0 = Id, \\ M_{k+1} = M_k + \dfrac{y_k y_k^T}{y_k^T s_k} - \dfrac{M_k s_k s_k^T M_k}{s_k^T M_k s_k}. \end{cases}$$

Recherche linéaire :
Le but de la recherche linéaire est de trouver un pas t_k (le long de la direction de descente d_k) qui permet de diminuer suffisamment la fonction f. Etant donné que la recherche linéaire est une boucle exécutée à chaque itération, il est donc essentiel qu'elle soit très rapide. Pour cela les recherches linéaires utilisées ne sont pas exactes. Elles cherchent dans un intervalle $[t_g, t_d]$ une valeur t_k jugée acceptable en utilisant le schéma suivant :

- `Etape 0` : t initial > 0 (initialisation par la boucle extérieure)
  $t_g = 0$ et $t_d = \infty$
- `Etape 1` : test de t
  - `(a)` t *acceptable*. Stop
  - `(b)` t trop grand, faire $t_d = t$ et aller en `Etape 2`
  - `(c)` t trop petit, faire $t_g = t$ et aller en `Etape 2`
- `Etape 2` : choix d'un nouveau t
  - si aucun $t_d \neq \infty$ n'a été trouvé : prendre un t plus grand (*extrapolation*)
  - si $t_d \neq \infty$: prendre $t \in]t_g, t_d[$ (*interpolation*)
- `Etape 3` : aller en `Etape 1`

L'extrapolation se fait en prenant par exemple, $t = a * t$ avec $a > 1$ et l'interpolation se fait en prenant $t = \frac{t_g + t_d}{2}$. Dans l'`Etape 1 :`, le test t *acceptable*, se fait à l'aide d'une règle garantissant la convergence de l'algorithme. Les règles les plus connues sont la règle d'Armijo, la règle de Wolfe et la règle de Goldstein et Price données ci dessous. On note que la règle d'Armijo n'est pas très efficace car elle n'exclut pas de prendre des pas "trop petits" ce qui ralentit l'algorithme. La Figure 1.7 illustre les différents cas pour la règle de Wolfe.

Règle d'Armijo

a) $\frac{q(t)-q(0)}{t} \leq m_1 q'(0)$

b) $m_1 q'(0) < \frac{q(t)-q(0)}{t}$

c) jamais !

Règle de Wolfe

a) $q(t) \leq q(0) + m_1 t q'(0)$ et $q'(t) \geq m_2 q'(0)$ t acceptable

b) $q(t) > q(0) + m_1 t q'(0)$ $t_d = t$

c) $q(t) \leq q(0) + m_1 t q'(0)$ et $q'(t) < m_2 q'(0)$ $t_g = t$

Règle de Goldstein et Price

a) $m_2 q'(0) \leq \frac{q(t)-q(0)}{t} \leq m_1 q'(0)$

b) $m_1 q'(0) < \frac{q(t)-q(0)}{t}$

c) $\frac{q(t)-q(0)}{t} < m_2 q'(0)$

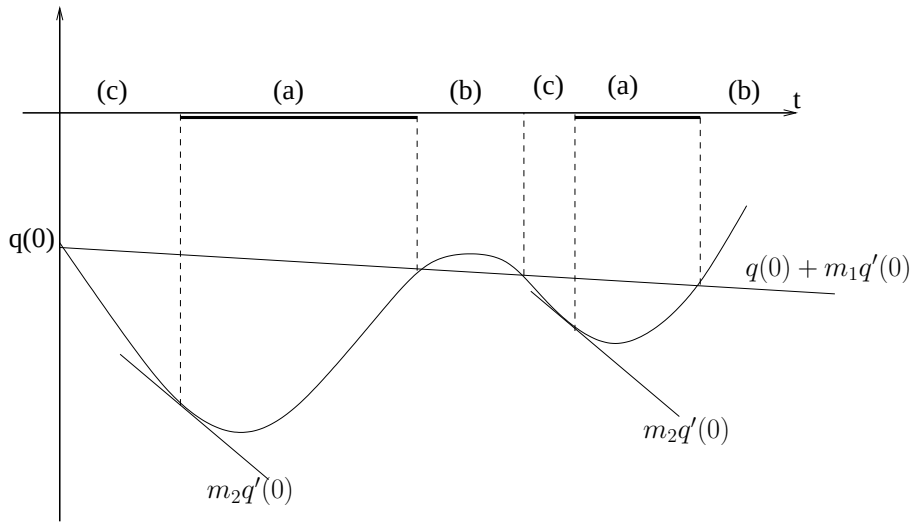


FIG. 1.7 – Schéma de la règle de Wolfe.

1.3.1.2 Calcul du gradient

Une première approche simple pour calculer le gradient d'une fonction $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$ est d'utiliser la méthode des différences finies :

$$\nabla J = \lim_{\delta x \to 0} \frac{J(x + \delta x) - J(x)}{\delta x}.$$

Cette méthode présente deux inconvénients majeurs. D'une part, elle est très sensible au pas de discrétisation et donc aux erreurs numériques. D'autre part, elle devient

très lente quand la taille de la variable x est grande car elle requiert $(n+1)$ évaluations de fonction.

La méthode de l'adjoint permet de s'affranchir de ces deux inconvénients. La méthode de l'adjoint peut être implémentée de plusieurs manières. On peut soit écrire les équations de l'état adjoint du problème continu et discrétiser les équations obtenues, soit discrétiser les équations directes et calculer l'état adjoint du problème discret. Dans le deuxième cas, la discrétisation peut se faire à deux niveaux. Une façon de faire consiste à discrétiser les équations en temps, à calculer l'état adjoint puis à discrétiser en espace (on parlera de semi-discrétisation). Une deuxième façon de faire consiste à calculer l'état adjoint des équations entièrement discrétisées en temps et en espace. On a testé ces différentes approches sur le problème de l'orientation afin de les comparer et de voir s'il y a une méthode meilleure que l'autre pour le calcul du gradient au moins dans le cas de notre problème. Les détails de cette étude se trouvent dans [P3, Section 3].

Les outils de différentiation automatique, apparus il y a une dizaine d'années, offrent une autre possibilité pour le calcul du gradient. Ils permettent de calculer le gradient à partir d'un programme informatique calculant la fonction $f(x)$. Le principe de ces outils est de "différencier ligne par ligne" les sources du programme pour construire le programme calculant ∇f. L'outil que l'on a utilisé est O∂yssée [13, 34] qui a évolué depuis en Tapenade [28]. Cet outil dispose de deux modes de différentiation, le mode direct, semblable au calcul du gradient par différences finies, et le mode adjoint, semblable à la méthode de l'adjoint. Dans la Section 1.3.1.3 on présente sur un exemple simple ces deux modes de différentation dans le cas d'O∂yssée .

Même si l'obtention du code calculant le gradient est rapide (quelques secondes dans notre cas), son utilisation n'est pas immédiate. En effet un travail de post-traitement est nécessaire afin d'éliminer les variables temporaires inutiles dans certains cas et qui peuvent consommer beaucoup de mémoire vive. Dans notre exemple le code brut donné par O∂yssée ne pouvait pas s'exécuter sur la machine utilisée (Pentium II, 466 Mhz Celeron avec 128 Mb RAM sous Linux) (voir Tableau 1.2). Après la suppression des variables temporaires inutiles dans les parties linéaires du programme, on a réduit de moitié la mémoire nécessaire à l'exécution du programme. Ceci était encore insuffisant et on a du supprimer le stockage de variables supplémentaires indispensables mais que l'on devait recalculer lorsqu'on devait les utiliser : on gagne en mémoire et on perd en temps de calcul.

1.3.1.3 Exemple d'un calcul de gradient par Différentiation Automatique

On présente dans cette section, sur un exemple simple, le principe de fonctionnement des modes de différentiation automatique. Pour calculer le gradient d'une

TAB. 1.2 – Données techniques du programme fourni par O∂yssée.

	code direct	code adjoint avant post-traitement	code adjoint après post-traitement
Taille (lignes)	433	2075	1190
Mémoire nécessaire	12 Ko	520 Mo	103 Mo
Temps (CPU)	60 s	—	141 s

fonction $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, on dispose de deux modes, le mode tangent (similaire aux différences finies) et le mode adjoint (similaire à la méthode de l'adjoint). Le premier est à utiliser lorsque $n \ll m$ et le second lorsque $n \gg m$.

Dans cet exemple, on utilise O∂yssée qui travaille sur des programmes écrit en Fortran. Considérons la fonction f définie par :

$$f(v_1, v_2) = e^{\frac{v_1^2 \sin(v_1+v_2^2)}{v_1+v_2^2}},$$

pour laquelle on veut calculer le gradient $\nabla f = (\frac{\partial f}{\partial v_1}, \frac{\partial f}{\partial v_2})$.

Le programme (sans les entêtes et les déclarations) calculant la valeur de la fonction est donné par :

```
ligne 1 :  v3 = v1 + v2**2
ligne 2 :  v4 = v1**2*sin(v3)
ligne 3 :  v4 = v4/v3
ligne 4 :  v5 = exp(v4)
```

Le code linéaire tangent calculant les dérivées directionnelles dans la direction $d = (d_1, d_2)$ est donné par :

```
v1tl = d1
v2tl = d2
v3tl = v1tl + 2*v2*v2tl
v3   = v1 + v2**2
v4tl = 2*v1*v1tl*sin(v3) + v1**2*cos(v3)*v3tl
v4   = v1**2*sin(v3)
aux  = v4/v3
v4tl = (v4tl - aux*v3tl)/v3
v4   = aux
aux  = exp(v4)
v5tl = aux*v4tl
v5   = aux
```

Le gradient s'obtient en appelant une première fois le programme précédent avec $(d_1, d_2) = (1, 0)$ pour calculer $\dfrac{\partial f}{\partial v_1}$. On appelle ensuite le programme précédent avec $(d_1, d_2) = (0, 1)$ pour calculer $\dfrac{\partial f}{\partial v_2}$. Le coût en temps de calcul pour le gradient est alors 2 fois (2 est le nombre de variables dans cet exemple, dans le cas général c'est n fois) celui de l'évaluation de la fonction f.

Le calcul du gradient en mode adjoint se fait en deux temps. Dans un premier temps, on calcule et on sauvegarde la trajectoire. Ce calcul consiste à refaire le calcul de la fonction f en y ajoutant des sauvegardes des états intermédiaires de certaines variables nécessaires dans la suite. Dans l'exemple considéré, cette étape est la suivante :

```
save3   = v3
v3      = v1 + v2**2
saveaux = aux
aux     = sin(v3)
save4   = v4
v4      = v1**2*aux
save41  = v4
v4      = v4/v3
save5   = v5
v5      = exp(v4)
```

Ensuite, le calcul en mode adjoint se fait en "remontant la trajectoire" (l'équivalent du calcul de l'état adjoint dans le cas de la méthode de l'adjoint). Dans notre cas on obtient le code suivant :

```
v1ad  = 0
v2ad  = 0
v3ad  = 0
v4ad  = 0
auxad = 0
v5ad  = 1
v5    = save5
v4ad  = v4ad + exp(v4)*v5ad
v5ad  = 0
v4    = save41
v3ad  = v3ad - v4ad*v4*v3ad/v3**2
v4ad  = v4ad/v3
v4    = save4
auxad = auxad + v4ad*2*v1*aux
v1ad  = v1ad + 2*v1*auxad
v4ad  = 0
aux   = saveaux
v3ad  = v3ad + auxad*cos(v3ad)
auxad = 0
v3    = save3
v1ad  = v1ad + v3ad
v2ad  = v2ad + v3ad*2*v2
v3ad  = 0
```

A la fin du calcul les composantes du gradient $(\frac{\partial f}{\partial v_1}, \frac{\partial f}{\partial v_2})$ sont stockées dans les variables $v1ad$ et $v2ad$. Le coût en temps de calcul pour l'obtention du gradient ne dépasse pas en général 5 fois celui de la fonction f, indépendamment du nombre de variables. Le coût en mémoire est, par contre, beaucoup plus important pour pouvoir sauvegarder la trajectoire. Ceci est d'autant plus pénalisant lorsque le programme différentié contient des boucles avec un grand nombre d'itérations. En effet, une variable qui est modifiée à l'intérieur d'une boucle nécessite pour sa sauvegarde un tableau dont la taille est le nombre d'itérations de la boucle. Il est donc primordial de repérer dans le programme les parties linéaires pour lesquelles il n'est pas nécessaire de sauvegarder la trajectoire pour calculer le gradient. Le gain en mémoire peut être non négligeable si ces parties se trouvent dans une longue boucle.

1.3.2 Algorithmes Évolutionnaires

1.3.2.1 Introduction

Le but d'un algorithme évolutionnaire est d'optimiser une fonction f dite *fonction objectif* sur un espace de recherche. Pour cela, une population d'individus, typiquement un P-uplet de points de l'espace de recherche, évolue selon un darwinisme artificiel (reproduction, mutation, sélection naturelle) basé sur la *fitness* F de chaque individu. La fitness est directement liée à la valeur de la fonction objectif f de cet individu (exemple, la fonction f elle-même). Des opérateurs appliqués à la population permettent de créer de nouveaux individus (croisement et mutation) et de sélectionner les individus de la population qui vont survivre (sélection et remplacement). Les opérateurs appliqués à un individu ne sont pas en général définis sur le même espace que celui sur lequel est défini la fonction fitness, appelé *espace des phénotypes*, mais sur un espace de *représentation* appelé *l'espace des génotypes*. Par exemple pour un codage binaire les algorithmes génétiques simples utilisent un espace de génotypes de la forme $\{0,1\}^n$. La Figure 1.8 illustre le schéma général d'un algorithme évolutionnaire : après l'initialisation de la population (généralement d'un façon aléatoire) l'algorithme évalue la fitness de chaque individu. La boucle de l'algorithme suit les étapes suivantes :

- *Critère d'arrêt :* un des critères simples souvent utilisé est lorsque le nombre maximum de générations, fixé par l'utilisateur, est atteint.
- *Sélection :* cet opérateur sélectionne parmi les parents ceux qui vont générer des enfants. Plusieurs opérateurs sont possibles qui peuvent être soit déterministes soit stochastiques. La sélection est basée sur la fitness des individus.
- *Création de nouveaux individus :* la création de nouveaux individus se fait essentiellement à l'aide des opérateurs de *croisement* et de *mutation*. L'opérateur de croisement est un opérateur stochastique qui combine k parents pour créer un ou plusieurs enfants. L'opérateur de mutation est un opérateur stochastique qui modifie un individu pour en créer un autre qui lui est généralement proche (ce qui dépend énormément de la représentation choisie).
- *Evaluation :* calcul de la fitness de chaque enfant. C'est l'étape la plus coûteuse en temps de calcul.
- *Remplacement :* on détermine qui parmi la population courante fera partie des parents de la génération suivante. Cet opérateur est basé, comme l'opérateur de sélection, sur la fitness des individus.

1.3.2.2 Les grandes familles d'Algorithmes Évolutionnaires

Tous ces algorithmes ont en commun de faire évoluer des populations d'individus. La différence entre eux est principalement d'ordre historique. On peut classer ces algorithmes en 4 grandes familles. Dans la suite, on détaille uniquement les deux

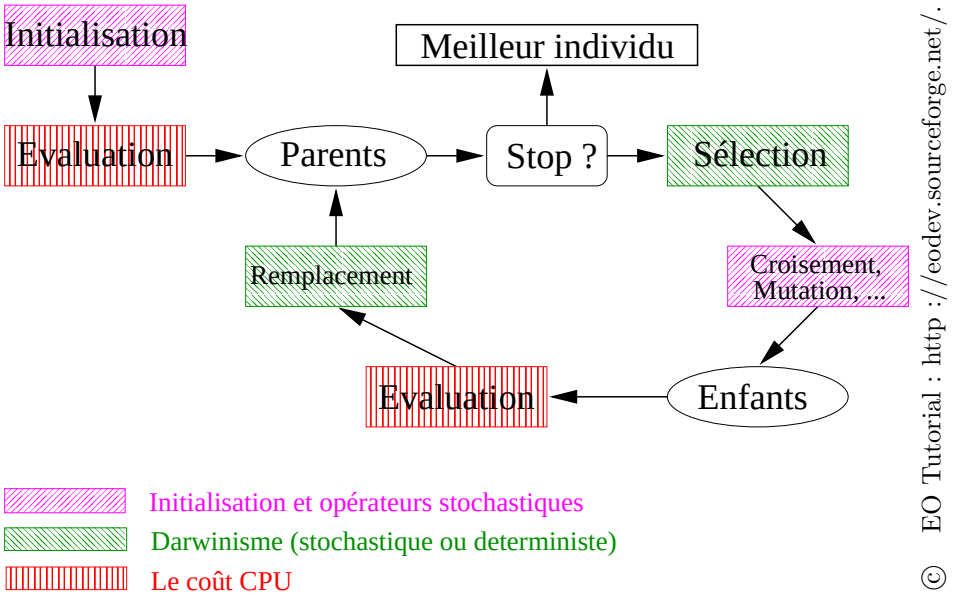


FIG. 1.8 – Schéma général d'un algorithme évolutionnaire.

familles d'algorithmes utilisés. Pour une description plus détaillée de ces algorithmes, on renvoie à [3, 23].

- Algorithmes Génétiques (AG) : J. Holland (1975) et D. E. Goldberg (1989).
  Les plus connus et les plus populaires des algorithmes évolutionnaires. Ils ont été développés pour modéliser l'adaptation des populations en biologie.
- Stratégies d'Evolution (ES) : I. Rechenberg et H. P. Schwefel (1965).
  Développés par des ingénieurs pour résoudre des problèmes d'optimisation paramétriques. Ces algorithmes sont les plus efficaces pour ce type de problèmes.
- Programmation Evolutionnaire (EP) : L. J. Fogel (1966).
  Développée à l'origine pour la découverte d'automates à états finis.
- Programmation Génétique (GP) : J. Koza (1990).
  Apparue initialement comme un sous-domaine des GAs, la programmation génétique est devenue une branche à part entière. La spécificité de ces algorithmes est de représenter des individus par des arbres.

Dans la suite, on présente les différentes mises en œuvre des algorithmes évolutionnaires. Dans la Section 1.3.2.3, on donne d'abord l'exemple d'un Algorithme Génétique Simple (AGS) utilisant des opérateurs basiques pour le croisement, la mutation et la sélection. Deux algorithmes évolutionnaires ont été utilisés : un algorithme génétique (AG), développé dans le cadre de ma thèse, et un algorithme de stratégies d'évolution (ES). L'algorithme ES, fournis par *EOlib class library* [12], a été utilisé dans le cadre d'un travail commun avec Auger [2]. Ces deux algorithmes AG et ES sont présentés respectivement dans la Section 1.3.2.4 et la Section 1.3.2.5. Dans la

Section 1.3.2.6 on présente un algorithme hybride développé à partir de l'algorithme AGE.

1.3.2.3 Algorithmes Génétiques Simples (AGS)

Les premiers AGs utilisaient un codage binaire avec un espace de génotype de la forme $\{0,1\}^n$. Les opérateurs de sélection testés sont la roulette et la roulette stochastique où la probabilité P_{X_p} de sélectionner un individu X_p est proportionnelle à sa fitness $F(X_p)$. Pour la roulette, P_{X_p} est donnée par :

$$P_{X_p} = \frac{F(X_p)}{\sum_{i \in Population} F(X_i)}.$$

L'opérateur de croisement le plus simple consiste à remplacer une partie des chromosomes de l'un des parents par ceux de l'autre parent. Par exemple, deux parents

$$X_1 = (x_1^1, x_1^2, \cdots, x_1^n) \text{ et } X_2 = (x_2^1, x_2^2, \cdots, x_2^n),$$

permettent de générer deux enfants

$$Y_1 = (x_1^1, x_1^2, \cdots, x_1^q, x_2^{q+1}, \cdots, x_2^n) \text{ et } Y_2 = (x_2^1, x_2^2, \cdots, x_2^q, x_1^{q+1}, \cdots, x_1^n),$$

où l'entier q est choisi aléatoirement dans $[1, n]$. L'un des point faibles du codage binaire est qu'un tel croisement peut "mélanger" des variables de type différent.

La mutation dans le cas d'un codage binaire consiste à changer un 0 par 1 ou inversement. Dans le cas d'un codage réel, la mutation peut se faire en remplaçant une variable x_i par $x_i + \delta x_i$ où δx_i est une "petite" variation de la variable x_i. Le remplacement se fait par un remplacement *générationel* : les enfants d'une génération n deviennent les parents de la génération $n+1$.

1.3.2.4 Algorithme Génétique utilisé (AG)

L'algorithme développé pour résoudre le problème de l'orientation a été écrit en Fortran 77 [33]. Cet algorithme utilise un codage réel. La sélection utilisée est soit la sélection par roulette soit la sélection par roulette stochastique. Les croisements disponibles sont le croisement multi-points et le croisement barycentrique. Dans le cas d'un croisement multi-points avec deux points, le croisement se fait de la manière suivante : deux parents

$$X_1 = (x_1^1, x_1^2, \cdots, x_1^n) \text{ et } X_2 = (x_2^1, x_2^2, \cdots, x_2^n)$$

génèrent deux enfants

$$Y_1 = (x_1^1, x_1^2, \cdots, x_1^{q_1}, x_2^{q_1+1}, \cdots, x_2^{q_2}, x_1^{q_2+1}, \cdots, x_1^n)$$

et

$$Y_2 = (x_2^1, x_2^2, \cdots, x_2^{q_1}, x_1^{q_1+1}, \cdots, x_1^{q_2}, x_2^{q_2+1}, \cdots, x_2^n),$$

où les entiers q_1 et q_2 sont choisis aléatoirement dans $[1, n]$. Avec un croisement barycentrique les enfants créés sont donnés par $Y_1 = (y_1^1, \cdots, y_1^n)$ et $Y_2 = (y_2^1, \cdots, y_2^n)$ avec pour tout i, $y_1^i = \alpha y_1^i + (1-\alpha) y_2^i$ et $y_2^i = \alpha y_2^i + (1-\alpha) y_1^i$ où α est un réel dans $[0, 1]$. Ce nombre α est soit choisi par l'utilisateur, soit choisi aléatoirement dans le cas d'un croisement barycentrique aléatoire.

Deux types de mutations sont possibles. La première est la mutation gaussienne avec une variance constante ou une variance décroissante au cours des itérations. La seconde est une mutation non-uniforme [23, 27] où une variable $x_i \in [x_{min_i}, x_{max_i}]$ prend la nouvelle valeur x_i' :

$$x_i' = \begin{cases} x_i + \Delta(t, x_{max_i} - x_i) & \text{si} \quad s \leq 0.5, \\ x_i - \Delta(t, x_i - x_{min_i}) & \text{si} \quad s \geq 0.5, \end{cases} \tag{1.9}$$

où t représente le nombre de génération, où s un nombre aléatoire dans $[0, 1]$ et où la fonction $\Delta(t, x)$ définie comme suit :

$$\Delta(t, x) = y \cdot r \cdot (1 - \frac{t}{T})^b,$$

avec r un nombre aléatoire dans $[0, 1]$, T le nombre maximal de générations et b est un paramètre de raffinement. L'allure de la fonction $\Delta(t, x)$ est représentée sur la figure 1.9. Ainsi, au début l'amplitude maximale de la mutation est grande alors qu'elle très petite vers la fin de l'algorithme : on passe d'une recherche globale à une recherche locale au cours de l'algorithme. Le paramètre b permet d'ajuster l'amplitude de la mutation. Le choix de b détermine la stratégie de compromis entre l'exploration et l'exploitation. Pour une grande valeur de b (Figure 1.10 (a)) on favorise l'exploration de l'espace et pour une petite valeur de b (Figure 1.10 (b)) on favorise la phase d'exploitation et de recherche locale.

L'algorithme utilise aussi la mutation dépendant de la distance (MDD) proposée dans [27]. L'idée est que deux parents qui sont proches vont générer des enfants qui "ressemblent" à leur parents, ce qui va limiter la diversité de la population et favorise l'apparition d'un super-individu qui peut attirer le reste de la population vers lui et engendrer une convergence prématurée. La MDD augmente la probabilité de mutation des enfants lorsque la distance entre les parents est petite.

Afin de favoriser l'exploration de l'espace de recherche la technique de *nichage* (niching) est utilisée. Cette technique consiste à favoriser les individus qui sont éloignés des autres individus de la population (ils explorent de nouvelles zones de l'espace de recherche) en augmentant leur fitness. L'*élitisme* a également été utilisé par l'algorithme. Il assure la décroissance de la meilleure valeur trouvée au cours des générations. Ainsi, si aucun enfant généré n'améliore le meilleur résultat déjà trouvé, le meilleur parent est automatiquement sélectionné pour la génération suivante.

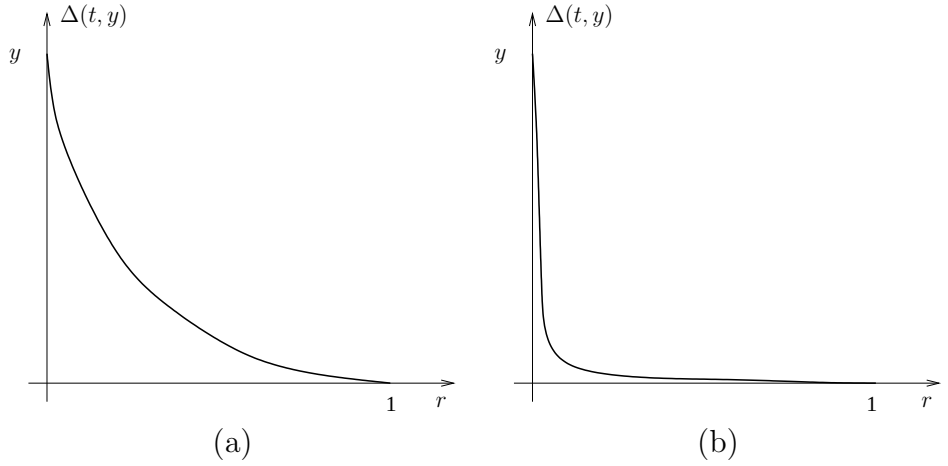


FIG. 1.9 – La fonction $\Delta(t,x)$ à deux instants t_1 et t_2 (resp. (a) et (b)) avec $t_1 < t_2$.

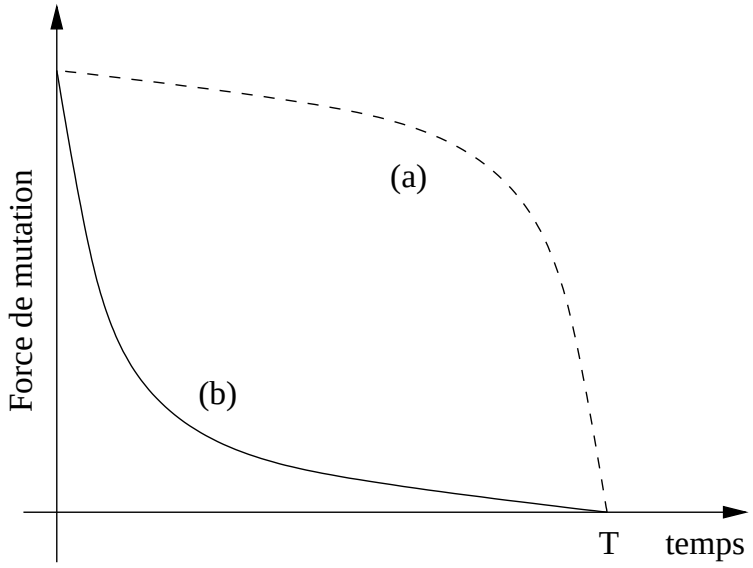


FIG. 1.10 – Évolution de l'amplitude de la mutation en fonction du temps.

Une autre technique, importante pour le bon fonctionnement de l'algorithme, est la mise à l'échelle (rescaling). Cette technique est imposée par le choix de la méthode de sélection basée sur la méthode de la roulette. Pour cette méthode, la probabilité P_{X_p} de sélectionner un individu X_p est proportionnelle à sa fitness $F(X_p)$:

$$P_{X_p} = \frac{F(X_p)}{\sum_{i \in Population} F(X_i)}.$$

Cette formule oblige à avoir une fitness positive pour tous les individus. D'où une première idée de scaling qui consiste à prendre pour fitness

$$F(X_p) = f(X_p) - \min_{i \in Population} f(X_i).$$

Ce choix est insuffisant car si on considère la fonction $g = f + C$ où C est une constante grande devant les valeurs de f, les individus ne sont plus réellement différents au vue de leur fitness donnée avec la fonction g. La sélection se fait alors différemment pour les fonctions f et g et devient sensible aux translations. Le choix du scaling doit satisfaire une autre condition dont le but est d'éviter qu'un super-individu (un individu très bon comparé au reste de la population) soit le seul sélectionné pour la génération suivante. Plusieurs stratégies de scaling satisfaisant à ces exigences ont été testées. La meilleure pour nos tests est le *sigma scaling tronqué :*

$$F(X_p) = f(X_p) + (\overline{f} - c \cdot \sigma),$$

où $\overline{f}$ et σ représentent respectivement la moyenne et la variance de la fonction f sur la population et où c est une constante généralement prise entre 1 et 5. Les valeurs négatives éventuelles de $F(X_p)$ sont tronquées et ramenées à zéro.

L'ensemble des techniques présentées ci-dessus ont permis d'accroître les performances de l'algorithme développé. Le désavantage est que le nombre de paramètres à ajuster a augmenté ce qui induit un réglage plus lent.

Le temps de calcul requis par l'algorithme est essentiellement dû au temps d'évaluation d'une fonction (voir Figure 1.8). Pour les problèmes étudiés (problème de l'orientation, génération d'harmoniques hautes (HHG),..) ces temps sont assez importants et varient entre quelques secondes et une dizaine de minutes. L'étape de l'évaluation de la fonction se fait indépendemment pour les différents individus de la population, ce qui la rend facilement parallelisable. Ainsi, une version parallèle de l'algorithme a été développée en utilisant le protocole PVM [16]. A chaque génération, l'évaluation de la fitness des nouveaux individus est répartie sur plusieurs processeurs. Le reste des tâches est géré par le programme principal (ou programme maître) sur un unique processeur. Ceci a permis un gain considérable en temps de calcul lors des différentes optimisations.

1.3.2.5 Stratégies d'Evolution (ES)

Les ES utilisent une représentation réelle des individus. La mutation utilisée est une mutation gaussienne de loi normale $\mathcal{N}(0, \sigma)$. La particularité des ES est que la variance σ fait partie du codage définissant un individu. Pour un ES isotrope (iso-ES) un individu est de la forme $I = (x_1, \ldots, x_N, \sigma)$. Il est de la forme $I = (x_1, \ldots, x_N, \sigma_1, \ldots, \sigma_N)$ pour un ES non isotrope (non-iso-ES). Plus récemment, des algorithmes CMA-ES [17] ont été développés, où cette fois-ci, on fait évoluer la matrice de corrélation de la variable aléatoire $\mathcal{N}(0, \sigma)$. Ainsi, les paramètres de la mutation subissent à leur tour les opérateurs de croisement et de mutation. On parle de *mutation adaptative* qui se déroule en deux étapes. D'abord, on mute les paramètre σ pour muter ensuite les variables x_i. Pour un iso-ES ces deux étapes sont :

$$\begin{aligned} \sigma^{(t+1)} &= \sigma^{(t)} \exp\left(\tau_0 \mathcal{N}(0,1) + \tau \mathcal{N}(0,1)\right), \\ x_i^{(t+1)} &= x_i^{(t)} + \mathcal{N}_i(0, \sigma^{(t+1)}). \end{aligned}$$

Et pour un non-iso-ES,

$$\begin{aligned} \sigma_i^{(t+1)} &= \sigma_i^{(t)} \exp\left(\tau_0 \mathcal{N}(0,1) + \tau \mathcal{N}_i(0,1)\right), \\ x_i^{(t+1)} &= x_i^{(t)} + \mathcal{N}_i(0, \sigma_i^{(t+1)}). \end{aligned}$$

L'opérateur de croisement sélectionne d'une manière aléatoire deux parents notés, $(x_1^1, \ldots, x_N^1, \sigma_1^1, \ldots, \sigma_N^1)$ et $(x_1^2, \ldots, x_N^2, \sigma_1^2, \ldots, \sigma_N^2)$, pour générer un nouvel enfant $(x_1^{q_1}, \ldots, x_N^{q_N}, \sigma_1^{q_1}, \ldots, \sigma_N^{q_N})$ avec $q_i = 1$ ou $q_i = 2$ d'une façon équiprobable. Ce croisement peut faire participer tous les parents, on parle alors de *croisement global*. L'opérateur de remplacement est strictement déterministe, basé sur le rang des individus. Par exemple l'algorithme dit $(\mu, \lambda) - ES$ sélectionne les μ parents de la génération suivante en prenant les μ meilleurs parmi les λ enfants créés. Pour l'algorithme $(\mu + \lambda) - ES$, on sélectionne les μ meilleurs individus parmi l'ensemble des parents et des enfants.

1.3.2.6 Algorithmes Hybrides (AG-GC)

Les méthodes hybrides combinent plusieurs méthodes d'optimisation afin d'améliorer leur efficacité à trouver l'optimum et d'accélérer leur convergence. L'approche la plus simple consiste à utiliser le résultat obtenu par une méthode stochastique comme point de départ d'un algorithme déterministe. Cette approche permet de compléter la recherche locale mais ne permet pas d'améliorer réellement le résultat final si l'algorithme stochastique a bien convergé.

Une autre approche, dite *basin hopping* [32], consiste à inclure à l'intérieur de la boucle d'optimisation stochastique une boucle d'optimisation déterministe. Plus précisément, l'algorithme stochastique utilisé est un *recuit simulé* qui minimise une

fonction $\widetilde{f}$ obtenue à partir de la fonction f en lançant en chaque point x un algorithme de gradient : $\widetilde{f}(x) = f(\overline{x})$ où $\overline{x}$ est le résultat de l'algorithme de gradient lancé à partir du point x. Ceci a pour effet de réduire les barrières de potentiel que l'algorithme de recuit simulé doit franchir pour passer d'un minimum local à un autre. La méthode est utilisée pour optimiser des problèmes assez difficiles comme celui de l'optimisation du potentiel Lennard-Jones. Une variante de cette méthode, qui utilise un algorithme génétique au lieu du recuit simulé, a été implémentée. La section 8.3.3.4 du chapitre 8 présente cette méthode ainsi que les résultats obtenus dans le cas de l'optimisation du potentiel Lennard-Jones.

L'approche adoptée dans l'algorithme AG-GC consiste à utiliser l'algorithme génétique, présenté dans la section précédente, en y ajoutant un opérateur de mutation par gradient. Cet opérateur remplace un individu x par le résultat de l'optimisation par un algorithme de gradient conjugué après un certain nombre d'itérations N_{iter}. Le choix de l'individu x peut se faire en prenant le meilleur individu de la population ou en choisissant un individu d'une manière aléatoire. Le deuxième choix s'est avéré plus efficace dans la plupart des tests effectués. En effet, le meilleur individu attire autour de lui d'autres individus de la population et donc, au cours des itérations une forme de recherche locale s'effectue déjà dans sa région. On note aussi que le nombre d'itérations N_{iter} ne doit pas être très grand pour ne pas ralentir le programme. De plus, au début du processus lorsque les individus sont loin de l'optimum il est inutile de raffiner la recherche locale.

1.3.3 Résultats sur des fonctions tests

Dans cette section on présente les résultats des simulations obtenues par les méthodes déterministes et les algorithmes évolutionnaires sur quatre fonctions tests prises dans la littérature [27]. Ces exemples ont été choisis de façon à illustrer le fonctionnement des algorithmes présentés et à mettre en évidence les spécificités de chacun d'entre eux. Il est à noter que pour chaque méthode un effort particulier a été fourni pour ajuster les paramètres utilisés. Ceci a été utile dans la suite pour l'ajustement des paramètres sur le problème de l'orientation.

Pour chaque cas test, on présente la moyenne sur 50 exécutions de la distance à l'optimum en fonction du nombre d'évaluations de la fonction. On a choisi trois fonctions unimodales, c'est à dire avec un seul minimum, et une fonction multimodale. Pour les trois premières fonctions on présente le comportement des algorithmes de gradient et BFGS. Et pour chaque fonction, on présente les résultats de cinq algorithmes stochastiques : un algorithme génétique simple (AGS), un algorithme génétique développé (AG) sans l'opérateur de mutation par gradient, un AG avec un opérateur de mutation par gradient (AG-GC), un algorithme de stratégie d'évolution isotropique (iso-ES) et un algorithme de stratégie d'évolution non-isotropique (non-iso-ES).

	Fonction Sphère	Fonction Elliptique	Fonction de Rosenbrock
GCNL	3	180	299
BFGS	3	61	64

Tab. 1.3 – Nombre d'évaluations de fonctions avant convergence.

1.3.3.1 La fonction Sphère

En premier lieu, on considère la fonction Sphère :

$$f(x) = \sum_{i=1}^{30} x_i^2. \tag{1.10}$$

Ce cas est trivial pour les méthodes de type gradient, (voir Tableau 1.3), qui par construction trouvent la solution en une seule itération. En revanche, ce cas présente un intérêt à la fois théorique et numérique pour les algorithmes évolutionnaires. Sur la Figure 1.12 (a) on constate que les algorithmes iso-ES et AG d'une part, et l'algorithme AGS d'autre part, présentent deux comportements différents. Contrairement à l'algorithme AGS, les deux premiers sont capables d'augmenter la précision de la solution cherchée. Ceci est dû à l'auto-adaptation pour les ES et à la variation de la force de mutation en fonction du temps pour l'algorithme AG. On remarque qu'il est naturel d'utiliser ici iso-ES car la fonction sphère est symétrique.

Afin d'éviter d'avoir une population initiale symétrique par rapport à l'optimum, ce qui pourrait "biaiser" la recherche, on a testé les trois algorithmes évolutionnaires avec une population initiale prise dans l'intervalle $[10 - \epsilon, 10]^{30}$ avec $\epsilon = 10^{-15}$. Les résultats ont montré que les algorithmes AG et ES sont capables de sortir de cette boite et d'atteindre le minimum.

1.3.3.2 La fonction Elliptique

Cette variante de la fonction Sphère est telle que les variables contribuent d'une façon inégale à la valeur de la fonction :

$$f(x) = \sum_{i=1}^{30} 1.5^{i-1} x_i^2. \tag{1.11}$$

D'après le Tableau 1.3, on observe une nette différence entre l'algorithme de gradient conjugué (d'ordre 1) et l'algorithme BFGS (d'ordre 2). C'est un cas typique où les algorithmes d'ordre 2 type Newton ou Quasi-Newton sont beaucoup plus efficaces que le GCNL. Ceci est valable pour une dimension raisonnable N du problème. Pour une dimension très grande, les algorithmes du type Newton sont

ralentis considérablement du fait des opérations matricielles nécessaires au calcul de la direction de descente.

Étant donné que les variables contribuent différemment à la valeur de la fonction, l'algorithme iso-ES n'est plus adapté et c'est l'algorithme non-iso-ES qu'il faut utiliser. La Figure 1.12 (b), montre que ce dernier est plus efficace que l'algorithme AG qui ne tient pas compte la dissymétrie des rôles des variables. On note aussi que AGS est incapable de trouver l'optimum global (convergence prématurée). Par ailleurs, on observe que la mutation par gradient (AG-GC) accélère considérablement la convergence de l'algorithme AG.

1.3.3.3 La fonction de Rosenbrock

La fonction de Rosenbrock est donnée par :

$$f(x) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2, \tag{1.12}$$

et son optimum unique est le point $(1, 1)$. Elle présente une "large vallée" autour de ce minimum. Les algorithmes déterministes convergent assez rapidement vers l'optimum comme le montre le Tableau 1.3.

Pour les algorithmes évolutionnaires, cette fonction est un bon exemple où il faut trouver un équilibre entre la phase d'exploration et la phase d'exploitation : l'algorithme doit d'abord explorer la vallée et il doit ensuite converger localement vers l'optimum. Comme le montre la Figure 1.13 (a), contrairement à l'algorithme AGS, les algorithmes non-iso-ES et AG sont capables d'améliorer continuellement leur performance. On peut noter également, en observant la courbe de l'algorithme AG-GC, que la mutation par gradient accélère la convergence dans un premier temps. Par la suite, quand toute la population se trouve dans la partie plate de la vallée, le gradient est très proche de zéro. Ainsi, l'opérateur de mutation par gradient ne fait que ralentir la convergence au lieu de l'accélérer.

1.3.3.4 La fonction de Shekel

Cette fonction est un exemple typique où les algorithmes déterministes sont inefficaces. En dimension 2, la fonction de Shekel est définie par :

$$f(x) = 0.002 + \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^{2} (x_i - a_{ij})^6}, \tag{1.13}$$

et présente 25 minima dans le carré $[-64, 64]^2$. Les valeurs de la fonction en ces minima s'échelonnent entre 1 et 25 (voir Figure 1.11). A moins de partir d'un point initial proche d'un des minima (et dans ce cas, de converger vers ce minimum) les algorithmes de type gradient s'arrêtent dès la première itération, le gradient étant quasiment nul au point initial.

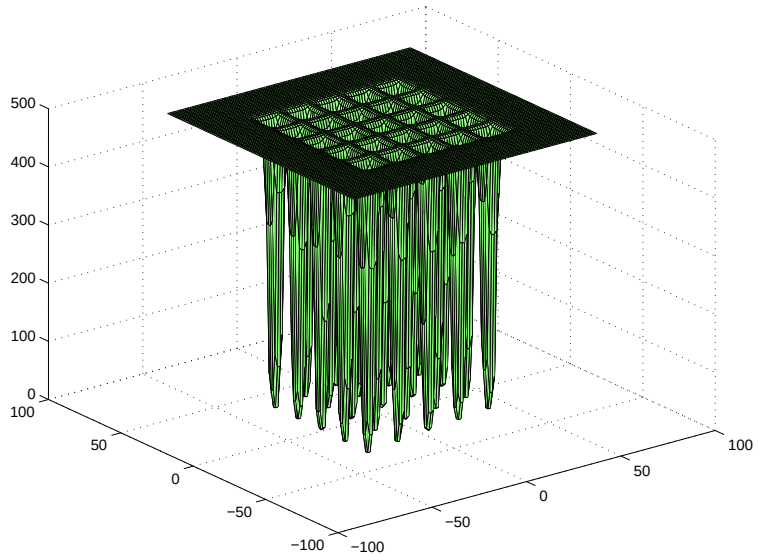
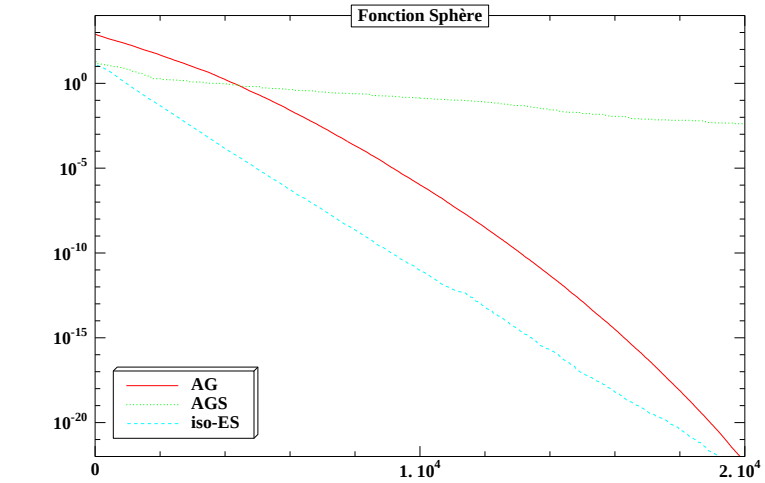


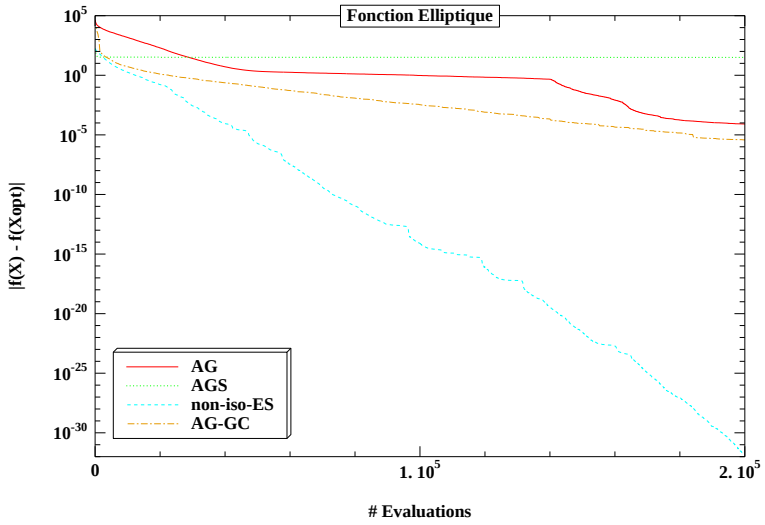
Fig. 1.11 – La fonction de Shekel.

Pour optimiser cette fonction un algorithme évolutionnaire doit avoir trois propriétés. Premièrement il doit être capable de trouver la bonne vallée contenant l'optimum global ; C'est la phase d'exploitation. Deuxièmement, il doit pouvoir y rester en gardant l'individu qui s'y trouve d'une génération à la génération suivante. Et troisièmement, il doit pouvoir effectuer efficacement la recherche locale dans cette vallée pour arriver jusqu'au minimum global. Sur la Figure 1.13 (b) on note que les algorithmes non-iso-ES et AG sont capables d'atteindre l'optimum global alors que AGS est incapable de trouver la bonne vallée. On observe également que la mutation par gradient n'améliore pas la convergence comme on pouvait le prévoir.

L'ensemble des résultats obtenus sur les cas tests, montrent que les algorithmes ES et AG sont plus performants qu'un simple AGS : ils convergent là où un simple AGS ne converge pas et donnent une meilleure précision en cas de convergence de ce dernier. Ils montrent également que l'utilisation de la mutation par gradient peut accélérer la convergence dans certains cas. Ces tests confirment aussi que les algorithmes déterministes et évolutionnaires ne sont pas en compétition mais qu'ils peuvent être complémentaires.

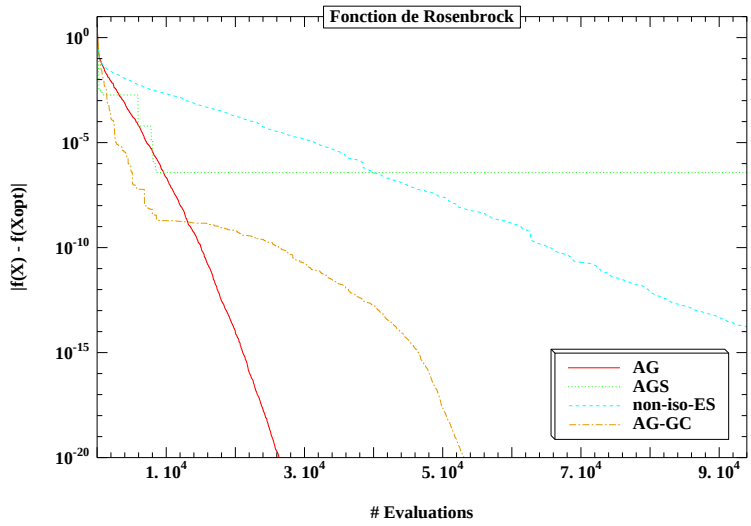


(a)

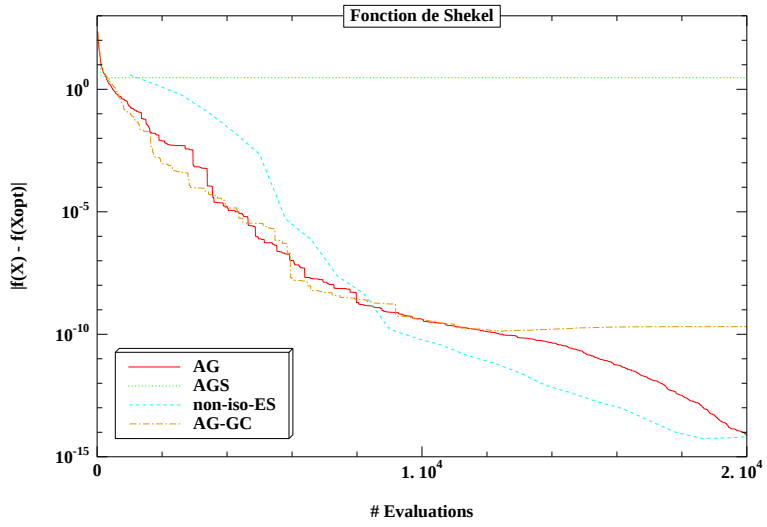


(b)

FIG. 1.12 – La fonction Sphère et la fonction Elliptique.



(a)



(b)

FIG. 1.13 – La fonction de Rosenbrock et la fonction de Shekel.

1.4 Les résultats obtenus sur le problème de l'orientation moléculaire

1.4.1 Le mécanisme de *kick*

L'un des résultats les plus intéressant obtenus est un résultat qui met en évidence un mécanisme pour orienter une molécule dit mécanisme de *kick*. Ce mécanisme a été déjà proposé en utilisant une impulsion en demie période (half-cycle pulses) [10]. Le champ laser ainsi trouvé permet de retrouver un mécanisme déjà connu en donnant une nouvelle façon de réaliser le champ laser correspondant. La nouveauté apportée par ce champ est qu'il a permis d'observer une orientation de la molécule en présence du champ laser (voir Figure 1.14).

Ce résultat a été obtenu en minimisant le critère $J_1 = j(T_{fin})$ avec la superposition de trois lasers et en utilisant la molécule HCN. L'algorithme d'optimisation (AG) a permis de trouver un champ laser qui est la superposition des trois champs dont les caractéristiques sont données par le Tableau 1.4. On observe sur ce tableau que l'un des trois champ est quasiment nul comparé aux deux autres champs ($\mathcal{I}_0 = 10^8 \ll \mathcal{I}_1 = \mathcal{I}_2 = 3 \times 10^{12}$). De plus ces deux derniers champs ont (aux décimales près) la même intensité, la même fréquence et un déphasage de π.

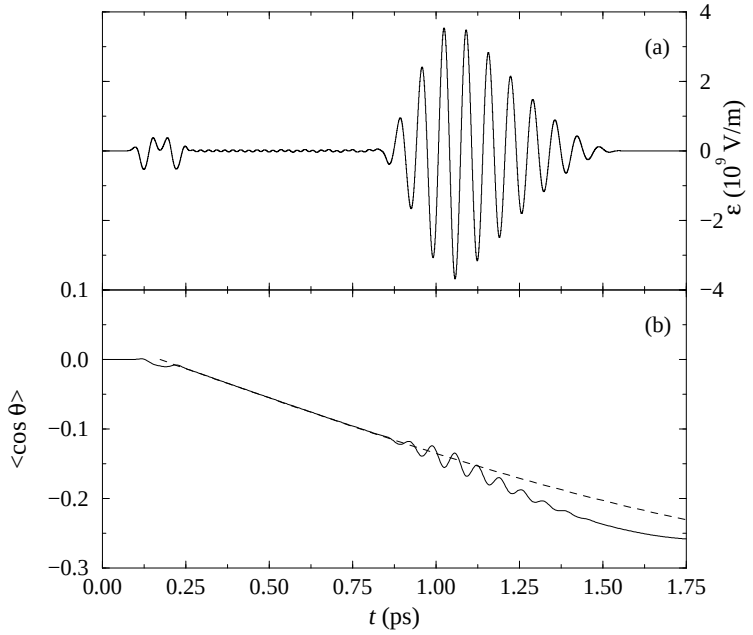


FIG. 1.14 – Le champ laser kick.

TAB. 1.4 – Paramètres du champ laser optimisé.

n	$\mathcal{I}_n$ (W/cm^2)	ω_n (cm^{-1})	ϕ_n (π rad)	t_{n0} (ps)	t_{n1} (ps)	t_{n2} (ps)	t_{n3} (ps)
1	1.01364×10^8	1389.541	1.98066	0.	0.312024	0.613023	1.193727
2	2.99976×10^{12}	500.051	1.82249	0.075077	0.270294	0.838110	1.562814
3	2.99989×10^{12}	500.000	0.82337	0.109518	0.235767	0.808280	1.080066

Ainsi, le champ laser trouvé est composé essentiellement de la superposition de deux champs laser qui sont identiques et en opposition de phases et qui ne diffèrent que par leur temps d'allumage et d'extinction. De plus, on observe sur la Figure 1.14 que la partie qui forme la fin du champ laser ne semble pas jouer un rôle dans l'orientation obtenue. Afin de vérifier ces hypothèses on a construit un champ laser à partir de deux champs ayant exactement les mêmes caractéristiques et en opposition de phases. Le temps d'allumage utilisé est celui donné par le champ optimisé et qui aussi servi pour construire le temps d'extinction. La Figure 1.15 montre que le champ laser ainsi construit oriente la molécule HCN de la même façon que le champ optimisé. Des résultats d'orientation semblables ont été reproduits avec ces deux champs (optimisé et reconstruit) sur la molécule LiF. L'idée du mécanisme de

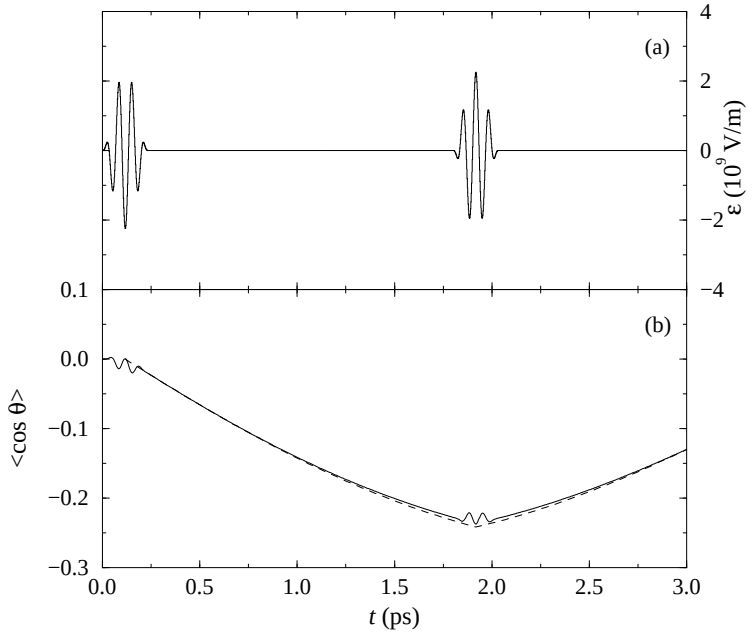


FIG. 1.15 – Le champ laser construit.

kick est renforcée en appliquant une *approximation soudaine* (sudden-impact) [10] qui donne une orientation semblable comme le montre les Figures 1.14 et 1.15. Cette approximation est valable car la durée de l'implusion (0.25 ps) est courte comparée à la période rotationnelle de la molécule (11.45 ps) :

$$t_{kf} - t_{ki} \ll \frac{\hbar}{B\hat{J}^2}.$$

Dans cette approximation, la soltuion de l'équation de Schrödinger est approchée par :

$$\psi(\theta, \varphi; t > t_k) = \exp\left(-\frac{i}{\hbar} B\hat{J}^2 t\right) \exp\left[i\left(\mathcal{A}\cos\theta + \mathcal{B}\cos^2\theta + \mathcal{C}\right)\right] \psi(\theta, \varphi; t = t_{ki}),$$

où $t_k = (t_{kf} - t_{ki})/2$, et où les cœfficients $\mathcal{A}$, $\mathcal{B}$, et $\mathcal{C}$ sont donnés par

$$\mathcal{A} = \frac{\mu_0}{\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}(t)dt,$$

$$\mathcal{B} = \frac{\alpha_\parallel - \alpha_\perp}{2\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}^2(t)dt,$$

et

$$\mathcal{C} = \frac{\alpha_\perp}{2\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}^2(t)dt.$$

Afin d'explorer le sous-ensemble de champs de type kick on a réalisé une optimisation du critère J_7 sur une famille de champs kick vérifiant : $\mathcal{I}_1 = \mathcal{I}_2 = 3.10^{13}$ W/cm^2, $w_1 = w_2 = 500$ cm^{-1} et $\phi_2 - \phi_1 = \pi$. Cette optimisation a permis de trouver un champ laser réalisant une orientation nettement meilleure que celle donnée par le précédent kick (voir Figure 1.16) aussi bien au niveau de l'intensité qu'au niveau de la durée.

1.4.2 Analyse des choix des critères

Dans cette partie on présente les principaux résultats d'une étude comparative des différents critères présentés dans la Section (1.2.3). Les détails de cette étude sont donnés dans [4]. L'étude a consisté à réaliser une optimisation pour chaque critère et à analyser les résultats ainsi produits.

Les critères utilisés au départ sont les critères J_1, J_2 et J_5 avec l'intervalle d'optimisation $[T_a, T_b] = [0, T_{fin}]$. Ceci a permis d'obtenir les premiers résultats d'orientation et en particulier le champ du type kick. A chaque fois, la durée de l'orientation obtenue est courte comparée à la période rotationnelle de la molécule T_{rot}. Afin d'obtenir une orientation sur des durées plus longues, on a élargi l'intervalle d'optimisation à $[T_a, T_b] = [T_{fin}, T_{fin} + T_{rot}]$. Cet intervalle est suffisant pour connaître

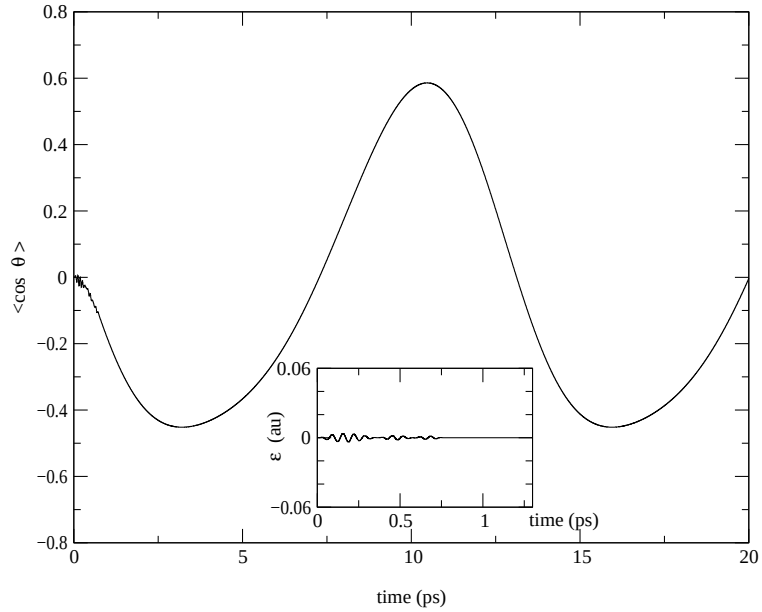


FIG. 1.16 – Un champ laser optimisé sur un sous-ensemble de champs de type kick.

l'évolution complète du système étant donné qu'en l'absence du champ laser, cette évolution est périodique de période T_{rot}. Ainsi les critères J_2 à J_7 ont été testés en utilisant l'algorithme génétique AG et en superposant 2 lasers.

L'optimisation du critère J_2 a produit une orientation intense ($\langle\cos\theta\rangle = 0.8$) mais de courte durée (.39 ps) étant donné que la durée n'est pas prise en compte par ce critère. Le dernier critère simple J_3 a produit un résultat sans intérêt car l'orientation obtenue est quasiment nulle pendant toute la période. Ce résultat était prévisible vu que ce critère ne prend pas en considération l'intensité de l'orientation. L'optimisation des critères hybrides a montré qu'il était possible d'optimiser à la fois l'intensité de l'orientation et sa durée. Le meilleur critère des quatre testés est le critère J_7 qui a donné une assez bonne orienation ($\langle\cos\theta\rangle = 0.68$) avec une durée de 0.7 ps.

Dans cette étude on a également réalisé des optimisations prenant en compte les effets de la température. Les critères J_2 et J_6 ont été optimisés en utilisant la moyenne thermique de l'orientation instantanée ($\langle j\rangle(t) = \langle\langle\cos\theta\rangle\rangle(t)$). On a obtenu des champs laser donnant une orientation assez robuste par rapport à la dispersion thermique ($\langle j\rangle(t) = \langle\langle\cos\theta\rangle\rangle = 0.38$ pour J_2 et 0.3 pour J_6 en utilisant 3 champs laser).

1.5 Perspectives

Afin de compléter cette étude sur le contrôle de l'orientation moléculaire plusieurs axes peuvent êtres explorés. D'abord, dans le cas de la dimension 1, on peut tester les approches multi-objectifs pour optimiser à la fois l'intensité et la durée de l'orientation. Cette approche permet de trouver des optima dits de Pareto pour lesquels on ne peut améliorer un critère sans détériorer l'autre. L'avantage d'une telle approche est que l'on n'ait pas à fixer une pondération a priori des critères. Ensuite, on peut améliorer la modélisation physique du problème en passant à un modèle bidimensionnel. Deux modèles physiques peuvent être ainsi envisagés. Dans le premier modèle, on considère l'équation (1.1) sans l'approximation du rotateur rigide. On prend en plus de l'angle θ, un degrés de liberté supplémentaire qui est l'une des distances inter-atomiques. Le deuxième modèle consiste à garder l'approximation du rotateur rigide et de considérer deux champs laser ayant deux axes de polarisation différents. On doit donc considérer la dépendance en la variable ϕ à cause de la perte de symétrie. Ce modèle est mieux adapté pour aborder le contrôle de l'orientation des molécules excitées sous l'effet de la température. En effet, une molécule ayant un état initial $\psi_0 = Y_{J,M}$ avec J et M élevés ne peut s'orienter avec un seul laser polarisé linéairement ce qui limite la valeur de la moyenne thermique de l'orientation qui peut être obtenue dans ce cas.

Les méthodes utilisées doivent être améliorées pour pouvoir traiter des problèmes plus complexes comme ceux que l'on obtient dans le cas du passage à la dimension 2. Les évaluations de la fonction de coût seront beaucoup plus coûteuses en temps de calcul dans ce cas. Les méthodes hybrides et l'exploitation des calculs en dimension 1 peuvent être un point de départ pour construire de nouvelles méthodes plus rapides et plus efficaces.

Bibliographie

[1] A. Assion, T. Baumer, M. Bergt, T. Brixnerand B. Kiefer, V. Seyfried, M. Strehle, and G. Gerber. Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses. *Science*, 282 :919–922, 1998.

[2] Anne Auger. *Algorithmes évolutionnaires pour la chimie*. PhD thesis, Université Paris 6, en préparation.

[3] Th. Bäck, D.B. Fogel, and Z. Michalewicz, editors. *Handbook of Evolutionary Computation*. Oxford University Press, Institute of Physics Publishing and Oxford University Press : Bristol and New York, 1997.

[4] J.M. Ball, J.E. Marsden, and M. Slemrod. Controllability for distributed bilinear systems. *SIAM J. Control Optim.*, 20 :575–597, 1982.

[5] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation Numérique : aspects théoriques et pratiques*. Springer, Berlin, 1997.

[6] E. Cancès, C. Le Bris, and M. Pilot. Contrôle optimal bilinéaire sur une équation de Schrödinger. *Note aux Compte Rendu de l'Académie des Sciences*, pages 567–571, 2000.

[7] C. E. Dateo and H. Metiu. Numerical solution of the time-dependent schrodinger equation in sperical coordinates by fourier-transformation methods. *J. Chem. Phys*, 95 :7392–7400, 1991.

[8] C. M. Dion. *Dynamique de l'alignement et de l'orientation moleculaire induite par laser. Simulations numeriques sur HCN en champ infrarouge.* PhD thesis, Université de Sherbrooke et Université de Paris-Sud, 1999.

[9] C. M. Dion, A. D. Bandrauk, O. Atabek, A. Keller, H. Umeda, and Y. Fujimura. Two-frequency ir laser orientation of polar molecules. numerical simulations for hcn. *Chem. Phys. Lett.*, 302 :215–223, 1999.

[10] C. M. Dion, A. Keller, and O. Atabek. Orienting molecules using half-cycle pulses. *Eur. Phys. J. D*, 14 :249–255, 2001.

[11] C. M. Dion, A. Keller, O. Atabek, and A. D. Bandrauk. Laser-induced alignment dynamics of HCN : Roles of the permanent dipole moment and the polarizability. *Phys. Rev. A*, 59 :1382–1391, 1999.

[12] EO. C++ class library, http ://eodev.sourceforge.net/.

[13] C. Faure and Y. Papegay. O∂*y*ssée User's Guide Version 1.7. *Rapport Technique INRIA RT-0224*, 1998.

[14] M.D. Feit, J.A. Fleck, and A. Steiger. Solution of a Schrodinger equation by a spectral method. *J. Comput. Phys.*, 47 :412–433, 1982.

[15] B. Freidrich and D. R. Herschbach. On the possibility of orienting rotationally cooled polar molecules in an electric field. *Z. Phys. D*, 18 :153–161, 1991.

[16] Al Geist, Adam Beguelin, Jack Dongarra, Weicheng Jiang, Robert Manchek, and Vaidyalingam S. Sunderam. *PVM : Parallel Virtual Machine : A Users' Guide and Tutorial for Networked Parallel Computing*, 1994.

[17] N. Hansen and A. Ostermeier. Convergence properties of evolution strategies with the derandomized covariance matrix adaptation : the $(\mu/\mu_I, \lambda)$-CMA-ES. In *Proceedings of the 5th European congress on intelligent techniques and soft computing (EUFIT'97)*, pages 650–654, 1997.

[18] R. S. Judson and H. Rabitz. Teaching lasers to control molecules. *Phys. Rev. Lett.*, 68 :1500–1503, 1992.

[19] J. J. Larsen, H. Sakai, C. P. Safvan, I. Wendt-Larsen, and H. Stapelfeldt. Aligning molecules with intense nonresonant laser field. *J. Chem. Phys.*, 111 :7774–7781, 1999.

[20] J. J. Larsen, I. Wendt-Larsen, and H. Stapelfeldt. Controlling the branching ratio of photodissociation using aligned molecules. *Phys. Rev. Lett.*, 83 :1123–1126, 1999.

[21] Mette Machholm. Postpulse alignment of molecules robust to thermal averaging. *J. Chem. Phys.*, 115 :10724–10730, 2001.

[22] Mette Machholm and Niels E. Henriksen. Field-free orientation of molecules. *Phys. Rev. Lett.*, 87 :193001, 2001.

[23] Z. Michalewicz. *Genetic algorithms + data structure = evolution programs.* Springer, 1999.

[24] R. Numico, A. Keller, and O. Atabek. Laser-induced molecular alignment in dissociation dynamics. *Phys. Rev. A*, 52 :1298–1309, 1995.

[25] Juan Ortigoso, Mirta Rodríguez, Manish Gupta, and Bretislav Friedrich. Time evolution of pendular states created by the interaction of molecular polarizability with a pulsed nonresonant laser field. *J. Chem. Phys.*, 110 :3870–3875, 1999.

[26] H. Sakai, C. P. Safvan, J. J. Larsen, K. M. Hilligsoe, K. Hald, and H. Stapelfeldt. Controlling the alignment of neutral molecules by a strong laser field. *J. Chem. Phys.*, 110 :10235–10238, 1999.

[27] Mourad Sefrioui. *Algorithmes Evolutionnaires pour le calcul scientifique. Application la mécanique des fluides et l'électromagnetisme.* PhD thesis, Université Pierre et Marie Curie, Avril 1998.

[28] Projet Tropics - INRIA Sophia-Antipolis. Tapenade : On-line automatic differentiation engine. http ://tapenade.inria.fr :8080/tapenade/.

[29] G. Turinici. Controlabilité exacte de la population des états propres dans les systèmes quantiques bilinéaires. *Note aux Compte Rendu de l'Académie des Sciences*, pages 327–332, 2000.

[30] G. Turinici and H. Rabitz. Quantum wave function controllability. *Chem. Phys.*, 267 :1–9, 2001.

[31] G. Turinici and H. Rabitz. Wavefunction controllability in quantum systems. *Preprint*, 2001.

[32] D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A*, 101 :5111–5116, 1997.

[33] A. Ben Haj Yedder. MyGa : a Genetic Algorithm in Fortran., http ://cermics.enpc.fr/∼benhaj/MyGa/.

[34] A. Ben Haj Yedder, E. Cancès, and C. Le Bris. Optimal laser control of chemical reactions using automatic differentiation. In George Corliss, Christèle Faure, Andreas Griewank, Laurent Hascoët, and Uwe Naumann (eds.), editors, *Proceedings of Automatic Differentiation 2000 : From Simulation to Optimization*, pages 203–213, New York, 2001. Springer-Verlag.

Chapitre 2

Contrôl optimal de réactions chimiques utilisant la différentiation automatique

Ce chapitre est la reproduction d'un article paru dans *Proceedings of Automatic Differentiation 2000* [P1]. On y présente l'utilisation de la différentiation automatique avec l'outil O∂yssée . On donne dans ce chapitre les premiers résultats obtenus sur le porblème du contrôle par laser de l'orientation moléculaire.

Optimal Laser Control of Chemical Reactions Using AD

Adel Ben-Haj-Yedder, Eric Cances and Claude Le Bris

CERMICS, École Nationale des Ponts et Chaussées
6 & 8, avenue Blaise Pascal, Cité Descartes,
Champs sur Marne, 77455 Marne-La-Vallée Cedex 2, FRANCE

Abstract: This chapter presents an application of automatic differentiation to a control problem from computational quantum chemistry. The goal is to control the orientation of a linear molecule by using a designed laser pulse. In order to optimize the shape of the pulse we experiment with a nonlinear conjugate gradient algorithm as well as various stochastic procedures. Work in progress on robust control is also mentioned.

2.1 Introduction

We consider a linear triatomic molecule HCN [2] subjected to a laser field $\overrightarrow{\mathcal{E}(t)}$. Our purpose is to use the laser as a control of the molecular evolution (see [5] for the general background), and more precisely as a control of the *orientation* of the molecular system. On the basis of experiment, it is believed that controlling the alignment of a molecular system is a significant step towards controlling the chemical reaction the system experiences.

An isolated molecular system is governed by the time-dependent Schrödinger equation (TDSE)

$$\begin{cases} i\hbar\frac{\partial\psi}{\partial t}(t) = H\,\psi(t), \\ \psi(0) = \psi^0, \end{cases} \tag{2.1}$$

where $\psi(t)$ denotes the wave function of the molecule, and H is the Hamiltonian of the free molecular system. The Hamiltonian H can be written as $H = H_0 + V(x)$, where H_0 is a second order elliptic operator corresponding to the kinetic energy (typically $H_0 = -\Delta$), and $V(x)$ denotes a multiplicative operator accounting for the

potential to which the molecule is subjected. When a laser field $\overrightarrow{\mathcal{E}(t)}$ is turned on, the dynamics of the molecular system is governed by :

$$\begin{cases} i\hbar\frac{\partial\psi}{\partial t} = H\psi + \overrightarrow{\mathcal{E}(t)}\cdot\mathcal{D}(t)\,\psi, \\ \psi(0) = \psi^0, \end{cases} \tag{2.2}$$

where $\mathcal{D}(t)$ denotes the electric dipolar momentum operator. In our problem, $\mathcal{D}(t)$ is approximated at the second order by : $\mathcal{D}(t)\,\psi = (\overrightarrow{\mu_0} + \alpha\cdot\overrightarrow{\mathcal{E}(t)})\psi$, where $\overrightarrow{\mu_0}$ and α denote the permanent dipolar momentum and the polarisability tensor, respectively.

For the molecular system under study (the HCN molecule), the Hamiltonian can be written in a very convenient way by resorting to the Jacobi coordinates $(\mathbf{R},\theta,\varphi)$ (the notation $\mathbf{R}$ refers to the pair (R,r), see Figure 2.1)

$$H(\mathbf{R},\theta,\varphi,t) = T_{\mathbf{R}} + H_{rot}(\mathbf{R},\theta,\varphi) + V(\mathbf{R}) + H_{laser}(\mathbf{R},\theta,\varphi,t), \tag{2.3}$$

where

$$T_{\mathbf{R}} = -\frac{\hbar^2}{2\mu_{HCN}}\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{\partial}{\partial R}\right) - \frac{\hbar^2}{2\mu_{CN}}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right),$$

$H_{rot}(\mathbf{R},\theta,\varphi) =$

$$-\frac{\hbar^2}{2(\mu_{HCN}R^2+\mu_{CN}r^2)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right],$$

$H_{laser}(\mathbf{R},\theta,\varphi,t) = -\mu_0(R,r)\mathcal{E}(t)\cos\theta$

$$-\frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R,r)\cos^2\theta + \alpha_{\perp}(R,r)\sin^2\theta\right].$$

2.2 Models and Results

Our goal is to control the orientation of the molecular system with the laser beam direction. We optimize the objective functional (the criterion) J, a measure of the rate of orientation given by

$$J = \frac{1}{T}\int_0^T\left[\left(\int_0^{\frac{\pi}{2}} - \int_{\frac{\pi}{2}}^{\pi}\right)\mathcal{P}(\theta,t)\sin\theta\,d\theta\right]dt,$$

where $\mathcal{P}(\theta,t)$ is the angular distribution of the molecule given by :

$$\mathcal{P}(\theta,t) = \frac{\langle\Psi(\theta,t)|\Psi(\theta,t)\rangle_R}{\langle\Psi(t)|\Psi(t)\rangle}$$

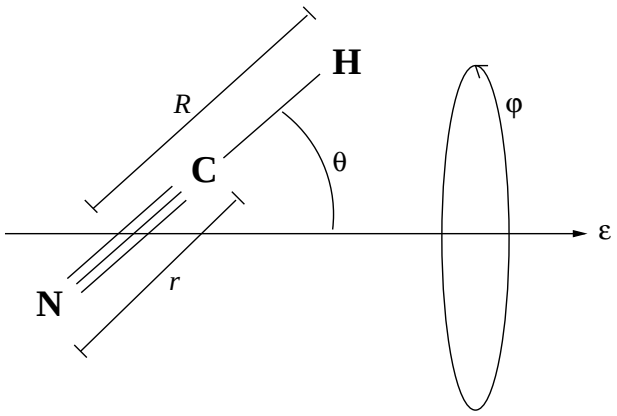


FIG. 2.1 – Model for HCN molecule

with

$$\langle f|g\rangle_R = \int_{r_{min}}^{r_{max}} \int_{R_{min}}^{R_{max}} f^*(R,r)g(R,r)R^2\,dR\,r^2\,dr,$$

and

$$\langle f|g\rangle = \int_0^{\pi} \int_{r_{min}}^{r_{max}} \int_{R_{min}}^{R_{max}} f^*(R,r,\theta)g(R,r,\theta)R^2\,dR\,r^2\,dr\sin\theta\,d\theta.$$

The criterion J takes its values in the range $[-1,1]$, the values -1 and 1 corresponding respectively to a molecule pointing in the desired direction and in the opposite direction. Our goal is therefore to minimize J. As a first step towards the treatment of the more sophisticated model (2.3), we consider here the case of a rigid rotor : the problem depends only on the variable θ. The equation (2.2) depending only on the variable θ is numerically solved by a Fortran program written by Dion [2] which uses a operator splitting method coupled with a FFT (for the kinetic part).

The laser field is the superposition of 10 laser pulses, each of the form

$$\mathcal{E}(t) = f(t)\mathcal{E}_0\cos(\omega t + \phi), \tag{2.4}$$

where :

$$f(t) = \begin{cases} 0 & \text{if} \quad 0 < t < t_0 \\ \sin^2\left[\frac{t-t_0}{t_1-t_0}\frac{\pi}{2}\right] & \text{if} \quad t_0 < t < t_1 \\ 1 & \text{if} \quad t_1 < t < t_2 \\ \sin^2\left[\frac{t-t_3}{t_2-t_3}\frac{\pi}{2}\right] & \text{if} \quad t_2 < t < t_3 \\ 0 & \text{if} \quad t > t_3\ . \end{cases} \tag{2.5}$$

Let us emphasize at this point that we do not pretend that such a superposition of laser beams is feasible experimentally : we are just testing here the mathematical attack of the problem.

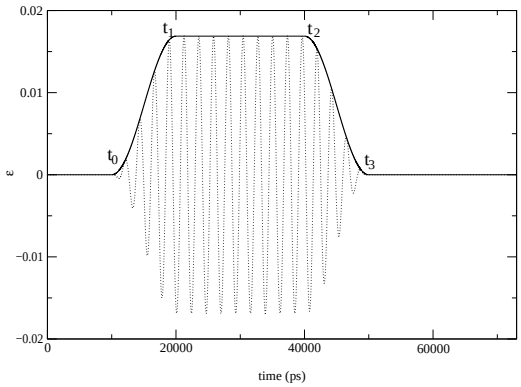


FIG. 2.2 – One laser field amplitude $\mathcal{E}_i(t)$

We first search for a local minimum of the criterion by means of a deterministic optimization algorithm (namely a nonlinear conjugate gradient procedure). The gradient is computed by an adjoint code automatically generated by O∂yssée [4]. In the present calculations, 70 parameters have to be optimized, namely $t_0^i, t_1^i, t_2^i, t_3^i, \mathcal{E}^i, \omega^i, \phi^i$ for $i = 1, 10$. As in the direct program we have 50,000 iterations, the adjoint program needs a lot of memory to run. To reduce the size of memory needed, the adjoint program was modified by deleting the temporary variables in the linear parts of the program. Table 2.1 gives an idea of the size of the direct code and the adjoint code. The reduction of the memory we have obtained by optimising the generated code by hand is coherent with the reduction obtained by an automatic optimisation as shown in [3]. The calculations are done on a Pentium II, 466 Mhz Celeron with 128 Mb RAM and running with Linux.

Numerical results show that the gradient values are most important for the variables ω^i and ϕ^i (which can thus be considered as the most significant ones from a physical viewpoint). When running the optimization program with a sample of representative initial guesses, it appears that the program always converges after a few iterations toward a *local* minimum generally located in the neighborhood of the chosen initial guess. This observation leads us to also turn to stochastic algorithms for searching a global minimum (see §2.3 below). Two cases of initial guesses were investigated : the case of laser pulses that approximately have the same frequencies, and the case of laser pulses with significantly different frequencies. In the first case, despite the number of local minima, it is nevertheless possible to reach quite a satisfactory local minimum ($J = -0.67$) by starting from an initial

TAB. 2.1 – Some data about the program

	direct code	standard adjoint	optimised adjoint
Size (lines)	433	2075	1190
Memory needed	12 Ko	520 Mo	103 Mo
Time (CPU)	60s	—	141s

TAB. 2.2 – Laser field characteristics for a first starting point for $i = 1, 10$

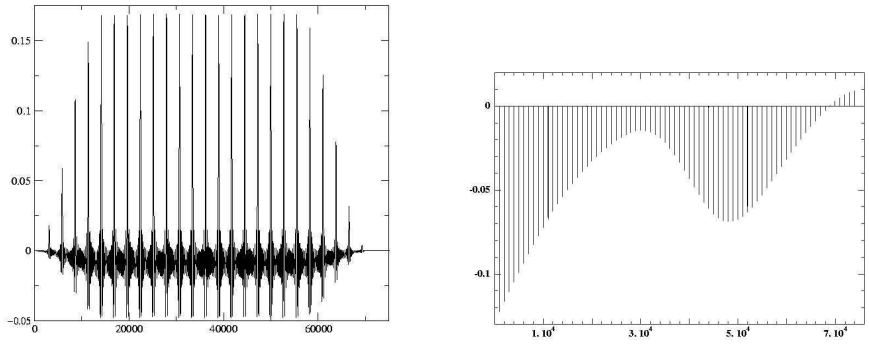
$\mathcal{E}$	ω	ϕ	t_0	t_1	t_2	t_3
10^{13}	1060	0	0	14000	56000	75000

guess where all the 10 laser pulses have the same characteristics (see Table 2.2). At this minimum, the values of the pulsations ω_i are very close to each other (for example $\omega_1 = 1060.52256$ and $\omega_2 = 1060.83702$). Unfortunately, a further analysis demonstrates that the so-obtained results are due to numerical instabilities : refining the discretization grid in both time and space makes the criterion go up to the value -0.0044 (a by far less good result !). With such close pulsations, a very fine grid is needed to avoid the numerical instabilities.

In the second case (see Table 2.3) the laser field found by the optimization program is stable with respect to a refinement of the grid. In this case, we have remarked that the laser field presents several very sharp peaks (see Figure 2.3 (a)). For this reason, consolidated by theoretical arguments, we have chosen next to perform optimization with laser pulses consisting not of functions of type (2.4) but rather in superpositions of Dirac functions. The Dirac functions are numerically approximated by Gaussian functions of the form : $\mathcal{E}_i \frac{1}{\sqrt{\delta t \pi}} \exp(-\frac{(t-t_i)^2}{\delta t})$. After some modifications of the direct code and a new automatic differentiation, we have used the new optimization program to minimize the criterion. The parameters are the instants t_i when the Dirac functions appear and the intensities $\mathcal{E}_i$ of the peaks. The calculations show that the gradient of the criterion with respect to the parameters t_i is indeed very small, and that the main control parameters are the intensities $\mathcal{E}_i$. The shape of the laser field given by the optimization program depends also on the number of Dirac functions. For 75 Dirac functions, the criterion $J = -0.56$, and the corresponding pulse shape is given in Figure 2.3 (b).

TAB. 2.3 – Laser field characteristics for a second starting point for $i = 1, 10$

$\mathcal{E}$	ω	ϕ	t_0	t_1	t_2	t_3
10^{12}	$500 \times i$	0	0	14000	56000	75000



(a) Laser field of type (2.4). (b) Laser field for Dirac functions.

FIG. 2.3 – Laser field given by the optimization program.

2.3 Comparison with Stochastic Algorithms

As already mentioned, the problem under investigation has many *local minima*. Stochastic procedures therefore seem appropriate to search for a *global minimum* in an effective way.

Our first trial consisted of using a simulated annealing algorithm with different temperature programs, which yielded disappointing results. We next combined the simulated annealing procedure with a nonlinear conjugate gradient algorithm, but the improvements were small compared to the cost in CPU time.

Using a genetic algorithm program given by [1], we got results better than those given by simulated annealing (although slightly worst than those given by the nonlinear conjugate gradient algorithm). However, we expect that coupling a genetic algorithm with conjugate gradient techniques will improve the search for a global minimum. Further work in this direction is in progress.

2.4 Robustness

To avoid instabilities (small variations of the laser pulse leading to large variations of the criterion) a very fine grid in space and in time is needed. However, such

a fine grid dramatically increases the computational cost in both memory and CPU time resources and will not be tractable in practice. Another way to avoid theses instabilities is to consider the new criterion $J = J(u, w)$, where u represents the same 70 parameters as above and where w represents a disturbance of the parameters [6]. The robust control problem consists of finding the solution of the minimax problem :

$$\min_u \max_w J(u, w).$$

Apart from the numerical instabilities mentioned above, robust control also enables us to take into account uncertainties due to the model itself. We have slightly modified the direct program and then used automatic differentiation to obtain the adjoint code. With a saddle point search algorithm the local optimization was slightly improved. This work is in progress and will be reported on in a future publication.

2.5 Conclusion

The results obtained so far by automatic differentiation for the problem of laser control of molecular systems seem to us very promising. It clearly brings some new contribution to the tools used by the community of chemists on the subject. For the different cases we have studied so far the optimization program provides a laser pulse which corresponds to a local minimum. Our work is now continuing along two directions : (a) implement *robust* control and (b) improve the search for a *global minimum*.

The use of AD tools in our work allows us to obtain better results (smaller objective function values), but less robust results, than results obtained by using stochastic algorithms. The combination of the two methods should make the optimization even more efficient. When we treat the complete problem (depending on all the molecular degrees of freedom R, r and θ, see Figure 2.1), checkpointing methods are likely to be necessary.

In parallel with the numerical work presented here, some experimental work is in progress.

Acknowledgments:

Special thanks are due to C. Dion for his help in understanding the direct code and for making the necessary modifications before automatic differentiation. We thank the O∂yssée development team for putting this software in our disposition for this work. We also wish to thank M. Barrault for his help and for useful discussions on optimization algorithms.

References

[1] David L. Carroll. FORTRAN genetic algorithm driver, Mar. 1999. See *http://www.staff.uiuc.edu/∼carroll/ga.html*.

[2] Claude Dion. *Dynamique de L'alignement et de L'orientation Moleculair'e Induite Par Laser. Simulations Numériques sur HCN En Champ Infrarouge*. PhD thesis, Université de Sherbrooke et Université de Paris-Sud, 1999.

[3] Christèle Faure and Uwe Naumann. Minimizing the tape size. In George Corliss, Christèle Faure, Andreas Griewank, Laurent Hascoët, and Uwe Naumann, editors, *Automatic Differentiation : From Simulation to Optimization*, Computer and Information Science, chapter 34, pages 293–298. Springer, New York, 2001.

[4] Christèle Faure and Yves Papegay. Odyssée User's Guide. Version 1.7. Rapport technique RT–0224, INRIA, Sophia-Antipolis, France, Sept. 1998. See *http://www.inria.fr/RRRT/RT-0224.html*, and *http://www.inria.fr/safir/SAM/Odyssee/odyssee.html*.

[5] Claude Le Bris. Control theory applied to quantum chemistry : Some tracks. In *International Conference on Control of Systems Governed by PDEs (Nancy, March 1999)*, volume 8, pages 77–94. ESAIM PROC, 2000. See *http://cermics.enpc.fr/reports/CERMICS-99-174.ps.gz*.

[6] Hong Zhang and Herschel A. Rabitz. Robust control of quantum molecular systems in presence of disturbances and uncertainties. *Physical Review A*, 49(4) :2241–2254, 1994.

Chapitre 3

Optimal laser control of molecular systems : methodology and results

Ce chapitre est la reproduction d'un article paru dans *Mathematical Models and Methods in Applied Sciences* [P3]. Dans ce chapitre on présente la méthodologie suivie pour traiter le problème du contrôle par laser de l'orientation moléculaire. En plus de la description des différents algorithmes utilisés, on détaille dans ce chapitre le calcul du gradient et la comparaison des approches testées.

Optimal laser control of molecular systems : methodology and results

A. AUGER, A. BEN HAJ YEDDER, E. CANCES, and C. LE BRIS
CERMICS, École Nationale des Ponts et Chaussées
6 & 8, avenue Blaise Pascal, Cité Descartes,
Champs sur Marne, 77455 Marne-La-Vallée Cedex 2, FRANCE

C. M. DION, A. KELLER and O. ATABEK
Laboratoire de Photophysique Moléculaire
Laboratoire de Photophysique Moléculaire du CNRS,
Bâtiment 213, Campus d'Orsay, 91405 Orsay, FRANCE

We report on some mathematical and numerical work related to the control of the evolution of molecular systems using laser fields. More precisely, the control of the orientation of molecules is our goal. We treat this as an optimal control problem and optimize the laser field to be used experimentally by using both deterministic and stochastic algorithms. Comparisons between the different strategies are drawn. In particular, when gradients of the cost functional are used, the different ways for their computation are compared and analyzed.

3.1 Introduction

We wish to report on theoretical and numerical work devoted to the modeling of the control of chemical reactions by laser fields. The laser control of chemical reactions is indeed a very active field of laser physics, at the crossroads between quantum chemistry, quantum mechanics, and theoretical and experimental femtophysics. Manipulation of molecular systems using laser fields is today an experimental reality [1], provided one restricts his aims to reasonable goals, as will be seen below. This leads to a mostly unexplored field for mathematical analysis and numerical simulation. Numerical simulations can indeed efficiently complement the experimental strategy, both by explaining the deep nature of the phenomena involved and by optimizing the parameters to be used experimentally.

We present here the contributions of our team, which is composed both of mathematicians and physicists. The emphasis is here on the mathematical aspects and the numerical techniques. A companion article [2] focusing on the physical aspects

appears elsewhere. The most striking result of our work is given in [17].

Before we discuss the technicalities, let us briefly state in a rather formal way the problem we shall deal with. All details will be given in Section 3.2, and for pedagogical purposes we prefer to only give a vague setting in this explanatory survey.

The evolution of a molecular system subjected to a laser field $\vec{\mathcal{E}}$ is modeled by the time-dependent Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} = H_0\,\psi + \vec{\mathcal{E}}(t)\cdot\vec{D}(\vec{\mathcal{E}}(t))\,\psi, \tag{3.1}$$

complemented with the initial condition $\psi(t=0)=\psi_0$. In this equation, the wave function ψ is assumed to depend only on the coordinates of the various nuclei the molecular system is composed of. The presence of the electrons is accounted for through an effective potential acting on the nuclei, and contained in the Hamiltonian H_0 of the free system (when the laser is turned off). We denote by $\vec{D}(\vec{\mathcal{E}}(t))$ the dipole moment of the molecule in presence of an external electric field $\vec{\mathcal{E}}(t)$; at the first order perturbation theory, one can use the form $\vec{D}(\vec{\mathcal{E}}(t)) = \vec{\mu}_0 + \overline{\overline{\alpha}}\vec{\mathcal{E}}$. More sophisticated models would feature higher order expansion of $\vec{D}(\vec{\mathcal{E}}(t))$ interactions, still in the perturbation setting, or even a true dependence of the wave function ψ and the Hamiltonian H with respect to the coordinates of *all nuclei and electrons* of the molecular system. To the present day, the latter model is out of reach of numerical treatment.

In order to state an optimal control problem, we need, in addition to the direct equation (3.1) modeling the evolution of the system, to define a cost function. Minimizing this cost function will give a formal sense to the physical target we want to reach. In our work, we consider a linear molecular system and intend to orient it in the direction of the linearly polarized laser field. The cost function we adopt will therefore reflect this wish. Just to fix the ideas, let us mention an example of cost function in the simplest case when the state of the system ψ (solution to equation 3.1) is a function of time t and of the angle θ between the axis of the system and the direction of the laser field :

$$J(\mathcal{E}) = \frac{1}{T}\int_{t=0}^{t=T}\int_{\theta=0}^{\theta=\pi} |\psi(t,\theta)|^2\cos\theta\,\sin\theta\,d\theta\,dt. \tag{3.2}$$

The reason why we choose an orientation problem as our control problem, and consequently such an objective function will be made clear below. Other forms of the cost function will also be given later in this article.

The simple setting we have just indicated above suffices to now underline the peculiarities of the optimal control problem we have to tackle, with respect to other optimal control problems that the reader may have in mind and that come from more usual domains of the engineering world (aeronautics, ...). Let us now emphasize these peculiarities.

From the standpoint of the mathematical theory, this problem is *bilinear* (the control $\vec{\mathcal{E}}$ multiplies the state ψ) which at once puts the problem on a very high

level of mathematical difficulty. Indeed, the mathematical theoretical results on bilinear control are very rare. In infinite dimension, i.e., for the PDE (3.1), since the celebrated work by Ball and Slemrod [4], no real progress has been made, to the best of our knowledge. For the finite dimensional approximation of (3.1), there exist some results that can also be extended to the infinite dimensional case but that are not very easy to exploit (so far). We refer to the work of G. Turinici et al. [33–35] for some recent progress on the theory of exact controllability for systems such as those we deal with here. For the optimal control problem, *some* minor things can be done. We refer in particular to [10] where some of us have proven the existence of an optimal field in a very academic and simplified setting. We shall not elaborate any longer on these theorical aspects and now concentrate on more practical ones.

A noticeable peculiarity is the fact that, in most cases, the control $\vec{\mathcal{E}}$ is *distributed in time*, and not in space. It is not a crucial fact for the sequel (cases when $\vec{\mathcal{E}}$ depends both on time and space could be treated in the same fashion, however with slightly more tedious computations) but it is rather convenient and constitutes a very reasonable approximation in the case of small molecular systems such as atoms and small molecules. At the scale of such a system, the laser light is indeed seen as homogeneous in space. Such a distributed in time control is not that usual for a partial differential equation such as (3.1).

In addition, special attention must be paid to the fact that although our goal is to drive the system from one initial state to some other specific state through a controlled time-dependent evolution, the cost function we choose to formulate our mathematical problem is not a distance to a target state, but the mean value of an observable (a measure of the orientation of the molecular system with the field). We wish to comment a little bit further on this point. The ultimate goal of the manipulations we want to model is the control of chemical reactions. This means for instance making a system ABC split into $AB+C$ rather than into $A+BC$ (see [9] for an introduction to this problem). Succeeding in making a chemical reaction possible does not necessarily mean driving the initial state to the final one, but sometimes (and in fact most of the times) only succeeding in *preparing* the initial system in a good way so that afterwards the desired reaction spontaneously happens. In that respect, orienting a molecule in space is both a modest and sufficient goal. Once it is conveniently "geometrically" prepared, the goal is almost reached. Nature will do the rest of the job. In addition, there are today experimental evidences showing that aligning a molecule (orientation is one step forward alignment) with a laser *is* feasible, and constitutes a significant step that can be used to efficiently control reactions (see the groundbreaking work by H. Stapelfeldt [27, 30]). The problem of orientation is therefore a good problem to look at.

It is also enlightening to consider this problem from the practical standpoint. Let us first indicate some orders of magnitude. Typically, the space scale is that of a molecule, namely a few angströms (10^{-10} m), and the time scale is that of the vibration of a molecular bond, namely ten femtoseconds (10^{-14} s). The total time of simulation for equation (3.1) is thus typically the picosecond (10^{-12} s). There are two main consequences of this time scale. First, the control needs to be an *open-loop* control, since it is clear that one cannot update the field in real time with electronic

devices. In other words, the only system that can react as fast as the molecular system is precisely the system itself. The second consequence is that we must think of this problem in a completely different way from the way we think of usual control problems : we are here in a framework where we can do thousands of experiments within a minute (while we cannot launch a rocket thousands times). This ability to make many experiments has in turn two consequences. First, one can imagine, and it is indeed done, to couple the numerical search for the optimal field not with the numerical simulation of equation (3.1), but with the experiment itself [1, 25]. The experimental solution of (3.1) is indeed much faster than its resolution on a computer. There is here some matter of reflection for experts in scientific computing. Second, one of the major problems of this field is the tremendous amount of data that are at our disposal. A challenge is to find a way to exploit them in the optimization cycle. We shall not give in this article any definite answer to the questions and concerns raised above, but it is sound to keep in mind these points.

One must also know about the practical parameters for a laser field. One of us has presented in [6, 7] a rapid account of this point, and we refer to it, or to the comprehensive report [13] for more details[b]. Let us only say that a trade-off has to be made between the power of the laser, its time resolution, its repetition frequency, and also its price and its size. The laser fields we shall make use of have intensities in the range $[10^{12}, 10^{13}]$ W/cm^2, are able to have a risetime of the order of 10^{-14}s, and the light they create has frequency around 10^{14}Hz. A very peculiar feature appears here again. One can ask the question whether it is better to optimize upon *only* the fields that are today experimentally feasible or to consider all fields without taking into account any contemporary technological constraint. Both approaches may be useful. In particular, the second one may help in designing the lasers physicists do need for the next generation. In the present article, we mostly choose the first approach, taking explicitly into account the technical requirements. We shall however also explore the second one (see more on this point below when we optimize with ten laser fields).

The stage is now set. Let us say a few words on the methodology we choose for the search of the optimal laser field.

First and foremost, we must emphasize that the present study is far from being the first attempt to find numerically the optimal laser field. There exist many theoretical studies based upon the construction of small systems of ODEs approximating (3.1) so that the optimal (or exact) control problem can be treated explicitly "by hand". The leading experts of this approach, fundamentally based upon a deep knowledge of (or intuition of) the main mechanisms are P. Brumer, P. Schapiro and coworkers [8,9]. Other outstanding contributions, in particular on intense laser fields are due to A. Bandrauk [11]. On the other hand, the optimal control methodology in the sense applied mathematicians speak about it has already been thoroughly explored by physicists, in the first row of which stands H. Rabitz [24,38]. See also works by Fujimura [23], Sakai [26]. However, in all these contributions, the algorithms used

[b]This report (in French !) presents a broad overview of the domain, indicating current approaches, both theoretical and experimental, and gives trends for the future. Another useful reference in the same spirit is [32].

for the numerical search for the optimized field are seen as black boxes, and not as topics for research. Our own approach aims at complementing the work of these leading researchers in physics by exploring the capabilities of the most recent optimization tools, by comparing them to one another on the present problem, by drawing conclusions on the best tools to be used, and also, when possible, by improving the physical conclusions.

On the present problem, we shall investigate mainly the following issues, which are of general interest, but whose response may differ from one problem to another :

- use on this specific case of deterministic algorithms (gradient-like algorithms), of stochastic algorithms (genetic algorithms and evolutionary strategies) and of algorithms mixing the two approaches, such as genetic algorithms accelerated by mutation by gradient
- comparison of the different ways to compute the gradient when needed : discretization of the adjoint equation, computation of the adjoint of the discrete equation, automatic differentiation
- impact of the choice of the cost function on the result, multicriteria approaches,...

The sequel of this article is organized as follows. In the next section, we give a detailed presentation of the problem under study, making more precise the quantities (Hamiltonian H, state ψ, electric field $\vec{\mathcal{E}}$, dipole moment $\vec{D}$, cost function J) we have described above in a somewhat vague way. Section 3.3 describes the different optimization methods we shall make use of. For some of them, we shall need to compute the gradient of the cost function. In Section 3.3.1, we therefore make a numerical analysis to determine which strategy is the best one to compute this gradient. In Section 3.3.2 we give a short description of stochastic algorithms we employed. Section 3.4 then gives the results obtained for our problem with deterministic algorithms and with stochastic ones. Finally, in Section 3.5, we shall summarize our main results and indicate the directions of our current and future research.

3.2 Statement of the control problem

3.2.1 The system under study and the control problem

The molecular system we study is the linear HCN molecule (hydrogen cyanide). This molecule has been chosen because it is linear in its ground state and should stay so if the laser frequency is out of resonance with respect to the bending modes. Therefore it constitutes a perfect toy object for testing our methodology. We use the so-called Jacobi coordinates ($\mathbf{R} = (R, r), \theta, \varphi$) to parameterize the state of the molecule (see Figure 3.1). The free Hamiltonian H_0 can be written as $H_0 = H_{\mathbf{vib}}(\mathbf{R}) + H_{rot}(\mathbf{R}, \theta, \varphi) + V(\mathbf{R})$ and the dipole moment is written as $D(\mathcal{E}(t)) = -\mu_0(R, r)\cos\theta - \dfrac{\mathcal{E}(t)}{2}\left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta\right]$. Then the gene-

ral form for the Hamiltonian, given in [16], is

$$H(\mathbf{R},\theta,\varphi,t) = H_{\mathbf{vib}}(\mathbf{R}) + H_{rot}(\mathbf{R},\theta,\varphi) + V(\mathbf{R}) + H_{laser}(\mathbf{R},\theta,\varphi,t), \tag{3.3}$$

where $T_{\mathbf{rot}} + H_{rot}$ denotes the kinetic energy operator with

$$H_{\mathbf{vib}}(\mathbf{R}) = -\frac{\hbar^2}{2\mu_{HCN}}\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{\partial}{\partial R}\right) - \frac{\hbar^2}{2\mu_{CN}}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right),$$

$$H_{rot}(\mathbf{R},\theta,\varphi) = -\frac{\hbar^2}{2(\mu_{HCN}R^2+\mu_{CN}r^2)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right],$$

where $V(\mathbf{R})$ denotes the effective potential resulting from the electrostatic interaction between nuclei and electrons (in their ground state), while

$$\begin{aligned} H_{laser}(\mathbf{R},\theta,\varphi,t) &= \mathcal{E}(t)\cdot D(\mathcal{E}(t)) \\ &= -\mu_0(R,r)\mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R,r)\cos^2\theta + \alpha_{\perp}(R,r)\sin^2\theta\right] \end{aligned}$$

denotes the interaction between the molecule and the laser field. In the former

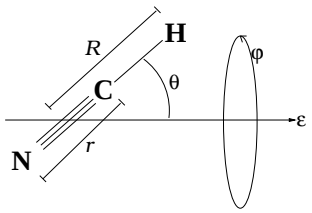


FIG. 3.1 – Model for the HCN molecule.

formulas, μ_{CN} and μ_{HCN} represent the reduced masses :

$$\mu_{CN} = \frac{m_C m_N}{m_C+m_N},\ \mu_{HCN} = \frac{m_H(m_C+m_N)}{m_H+m_C+m_N}$$

and μ_0 is the permanent dipole moment. The coefficients $\alpha_{\parallel}$ and $\alpha_{\perp}$ are respectively the parallel and the perpendicular components of the diagonal polarizability tensor $\overline{\overline{\alpha}}$ given by $\alpha_{\parallel} = \alpha_{zz}$ and $\alpha_{\perp} = \alpha_{xx} = \alpha_{yy}$ when (Oz) is the molecular axis.
As a first step toward the treatment of the sophisticated model (3.3), we consider in all the remainder of this article the case of a rigid rotor : the problem depends only on the angular variables θ, ϕ. Furthermore, symmetry conservation around the laser polarization axis allows us to separate the motion in ϕ from the motion in θ, and consider only the latter in our calculations. The Hamiltonian (3.3) therefore reduces to

$$H = H(\theta,t) = H_{rot}(\theta) + H_{laser}(\theta,t), \tag{3.4}$$

with

$$H_{rot}(\theta) = -\frac{\hbar^2}{2(\mu_{HCN}R^2 + \mu_{CN}r^2)}\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)$$

and

$$H_{laser}(\theta, t) = -\mu_0(R, r)\mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta\right],$$

where R and r are fixed at their equilibrium value. The objective function $J(\mathcal{E})$ we are optimizing will be detailed in Section 3.2.3 but let us now introduce the instantaneous criterion $j(t)$ used to compute $J(\mathcal{E})$ and which is the measure of the orientation at time t (see [21] for more details),

$$j(t) = \langle\cos\theta\rangle = \int_0^{\pi} \cos\theta \;\; \mathcal{P}(\theta, t)\sin\theta\, d\theta, \tag{3.5}$$

where $\mathcal{P}(\theta, t)$ is the angular distribution of the molecule. In the case of rigid rotor angular distribution is reduced to $\mathcal{P}(\theta, t) = \|\psi\|^2_{\mathbb{C}}$ where $\|\psi\|^2_{\mathbb{C}}$ denotes the squared norm of the complex ψ. The instantaneous criterion therefore becomes

$$j(t) = \int_0^{\pi} \cos\theta \;\; \|\psi\|^2_{\mathbb{C}}\sin\theta\, d\theta. \tag{3.6}$$

The instantaneous criterion $j(t)$ takes its values in the range $[-1, 1]$, the values -1 and 1 corresponding respectively to a molecule pointing in the direction of the laser field polarization axis and in the opposite direction.

The Schrödinger equation

$$\begin{cases} i\hbar\dfrac{\partial\psi}{\partial t} = H\,\psi, \\ \psi(t = 0) = \psi_0. \end{cases} \tag{3.7}$$

depending only on the variable θ is numerically solved with an operator splitting method [20] coupled with a FFT for the kinetic part as shown in [12, 29]. Table 3.1 summarizes the parameters of the HCN molecule for R and r fixed at their equilibrium value.

Tab. 3.1 – Parameters of the HCN molecule.

$B = \frac{\hbar^2}{2(\mu_{HCN}R^2 + \mu_{CN}r^2)}$ (a.u.)	μ_0 (a.u.)	$\alpha_{\parallel}$ (a.u.)	$\alpha_{\perp}$ (a.u.)
6.638×10^{-6}	1.141	20.05	8.638

3.2.2 Choice of the set of electric fields

We now describe the set of laser fields we minimize upon. As said in the introduction, both strategies of restricting oneself to the experimental state of the art or of considering the most general laser fields are of some interest. We begin with the second one, by considering that the electric field $\mathcal{E}(t)$ we have at our disposal is the sum of N ($\leq$ 10) individual linearly-polarized pulses : $\mathcal{E}(t) = \sum_{n=1}^{N} \mathcal{E}_n(t) \sin(\omega_n t + \phi_n)$. The envelope functions $\mathcal{E}_n(t)$ are of given sine-square form,

$$\mathcal{E}_n(t) = \begin{cases} 0 & \text{if} \quad t \leq t_{0n} \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t-t_{0n}}{t_{1n}-t_{0n}}\right)\right] & \text{if} \quad t_{0n} \leq t \leq t_{1n} \\ \mathcal{E}_{0n} & \text{if} \quad t_{1n} \leq t \leq t_{2n} \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t_{3n}-t}{t_{3n}-t_{2n}}\right)\right] & \text{if} \quad t_{2n} \leq t \leq t_{3n} \\ 0 & \text{if} \quad t \geq t_{3n} \end{cases} \tag{3.8}$$

each pulse being characterized by a set of 7 adjustable parameters, namely its frequency ω_n, relative phase ϕ_n, maximum field amplitude $\mathcal{E}_{0n}$, together with 4 times determining its shape (origin t_{0n}, rise time $t_{1n} - t_{0n}$, plateau $t_{2n} - t_{1n}$, and extinction time $t_{3n} - t_{2n}$). All beams are polarized along the same axis. This makes a total of

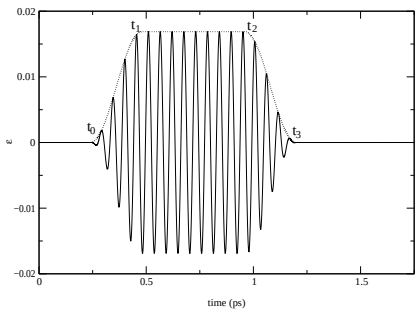


FIG. 3.2 – A typical laser field $\mathcal{E}_i(t)$.

$7 \times 10 = 70$ parameters. It should be once more emphasized that by considering such a superposition we do not have in mind to model a situation that is experimentally feasible, but only to generate a "generic" form of signal $\mathcal{E}(t)$.

As it will be seen below, using such a generic field has one main disadvantage (in addition to that obvious huge difficulty to minimize over $\mathbb{R}^{70}$!) : the optimized laser field that is obtained through minimization is likely to be too difficult to analyze ! Indeed, as we have very pragmatic purposes, we aim at providing the experimenter with a well identified field to generate. Obviously, a typical field obtained by such a minimization and shown on Figure 3.10 cannot be easily analyzed. Therefore,

the main part of our work will be along the first strategy : restrict ourselves to a superposition of two, or at most three, different lasers of the shape of Figure 3.2[c].

Apart from sticking to experimental reality (for instance a system of two lasers with the same pulsation but with two different phases is nothing else than the same laser with different optical paths), it greatly simplifies the post-treatment of results. In this view, one of our first results has been that when we use 3 lasers, i.e., when we allow for 3 different lasers in the minimization procedure, the algorithm ends up with an optimized field where the third laser has a very small amplitude (see Table 3.5 in Section 3.4). In other words, considering two lasers is enough. We shall therefore concentrate on this latter case.

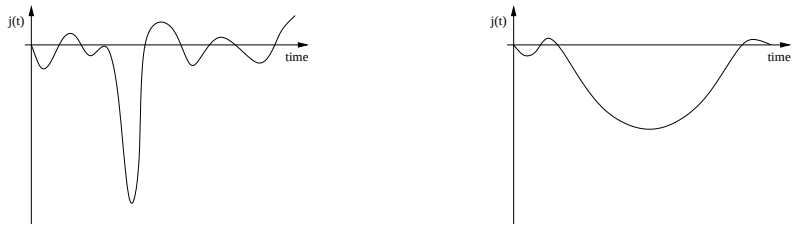
3.2.3 Choice of the cost function

The cost function is the mathematical formulation of our physical goal. Its choice is so difficult in our context that it has not been done *a priori*, but has been the result of an "iterative process". We have tested different ones and compared (on mathematical and physical bases) the results they produce. In this process, we have kept in mind the crucial following points : if a function produces (after minimization) a field which is too difficult to understand, it can be replaced however by another (possibly less) efficient that produces more understandable results. Most of the time we shall therefore handle many different cost functions, and not only one.
Basically, our physical goal is twofold :

- we want to have the molecule oriented with the field in a very good way at (at least) one time during the interval of time considered. The criterion for this purpose is :
$$J = \min_{t\in[0,T]} j(t), \tag{3.9}$$
- and/or we want this orientation to be kept as long as possible, even if it is not so perfect. Then the criterion to be used is :
$$J = \frac{1}{T}\int_0^T j(t)\,dt. \tag{3.10}$$

The latter criterion J is what we have written $J(\mathcal{E})$ in the equation (3.2). Unless otherwise mentionned, we shall deal henceforth with a criterion J that denotes either of the two criteria (3.9) or (3.10). In both formulas, let us recall that $j(t)$ is the quantity introduced in formula (3.6) of Section 3.2.1 and which the orientation at time t.
In the second setting, it should be made precise that "as long as possible" typically means relatively long compared to the rotation period of the molecule, namely 11 ps for HCN, which indeed is quite a long time in our context.
In the following, we shall call "narrow"(see Figure 3.3 (a)) a function $j(t)$ produced mainly by the optimization in the first setting and "wide" (see Figure 3.3 (b)) a function $j(t)$ produced mainly in the second one. Let us also mention that a

[c]even if the price to pay for this is to lose a little on the optimality



(a) (b)

FIG. 3.3 – Typical shape of an optimized $j(t)$ obtained with criterion J_1 (a) and criterion J_2 (b). They are respectively called a "narrow" and a "wide" in the text.

multicriteria approach is possible and that it may possibly result in obtaining many different minima and/or the best one in some sense to be defined (see Section 3.4).

3.2.4 Identification and classification of the fields obtained

Of primary interest is the need to understand the fields produced by the optimization algorithm. It will allow one to identify the underlying main mechanisms, to imagine scenarii, and to further simplify the electric field to suggest the most simple field to be experimentally generated.
The huge number of optimization processes we have run, with different sets of parameters, with different ranges of values of these parameters, and with different criteria, has resulted in an enormous data set of optimized fields $\mathcal{E}(t)$. We believe that a good way to classify them is :

– fields of the form of a *kick* (see Figure 3.4), which is an initial sudden (of approximately 0.25 ps, *i.e.*, much shorter than the rotational period of 11 ps) and asymmetric (with respect to its sign) pulse.

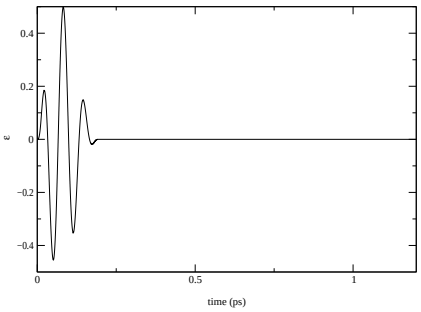


FIG. 3.4 – Example of a "kick" field.

– fields of the form $(\omega, 2\omega)$ (see Figure 3.5), which are a superposition of two laser fields with the pulsation of one being twice the pulsation of the other one.

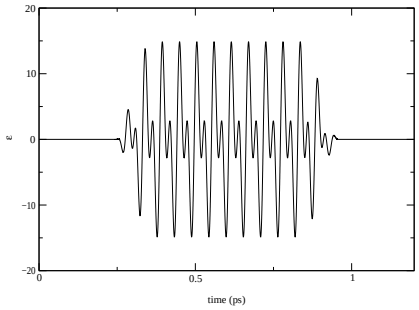


FIG. 3.5 – Example of an $(\omega, 2\omega)$ field.

– succession in time of two fields with a short overlay time (see Figure 3.6)

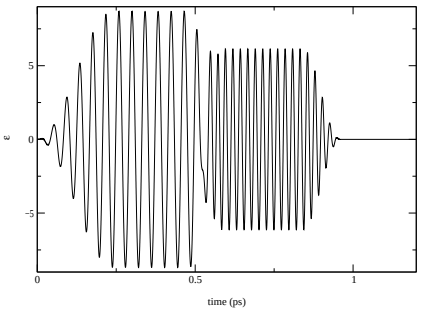


FIG. 3.6 – Example of a succession of two laser fields.

– other types of fields, apparently too complicated to be easily described.

3.3 Methodology

The way we have tackled the optimization of the orientation problem is based on two different classes of algorithms : first the gradient like algorithms, and second the evolutionary algorithms (EAs). The former ones are purely deterministic and are known to be from far the more rapidly convergent ones but present the drawback from running the risk of remaining trapped in a local minima. The latter ones are stochastic algorithms based on Artificial Darwinism. They are less sensitive to the number of local minima but as they are zero-order methods, the convergence is slower. A way to exploit the forces of both deterministic and stochastic algorithms is to use hybrid methods. We have explored one of these methods with an evolutionary algorithm using a gradient mutation operator.
This section presents in a first part different ways of computing the gradient of the criterion to optimize and compares the different methods. In a second part, this section briefly explains the basic steps of EAs, and next mention which purely stochastic and hybrid EAs have been used for the results presented in Section 3.4.

3.3.1 Gradient like algorithms

In this part, we present different ways to compute the gradient of the differentiable cost function $J(\mathcal{E})$, defined by (3.10) (the criterion (3.9) is not differentiable) needed for the gradient-like algorithms. We use two gradient like algorithms : the Polak-Ribière non linear conjugated gradient algorithm with Wolfe or Goldstein-Price line-search (hereafter abbreviate as PRLS) and the BFGS algorithm. For a complete presentation of these algorithms see [5, Part 1]. The most natural and the most easiest way to compute the gradient is the finite differences method, which is unfortunately very time consuming. So the need is to find another, less time consuming, way to compute the gradient. The well known adjoint method may be implemented in (at least) two ways : one can either discretize the continuous adjoint equation or one can do the adjoint calculus on the discretized form of the direct equation. It is not clear at all (at least to us) whether there is a general recipe claiming which of the two approaches is the best one. Therefore we shall test both approaches on our specific situation. In fact, the second approach (adjoint calculus on the discretized form) can be itself subdivided into two approaches : the semi-discrete approach, and the fully discrete one (see below). In addition, we shall also compare these methods with that of automatic differentiation (which in principle amounts to doing calculus on the fully discretized form of the equation, but which, in fact differs from this strategy because of implementation details). The tool we use in this latter approach is O∂yssée [19].

We begin in Section 3.3.1.1 by presenting the continuous approach which consists in discretizing the continuous adjoint equation. Next Section 3.3.1.2 details an intermediate approach where one does the adjoint calculus on the semi-discretized equations (which means equations only discretized in time) and next discretizes in space (θ) the so-obtained adjoint problem. In Section 3.3.1.3 we then compare this approach to the continuous one on a simplified example. In Section 3.3.1.4, we present the approach (called the discrete approach) consisting in doing adjoint calculus on the fully discretized equations (both in time and space). Finally, in Section 3.3.1.5, we present the automatic differentiation approach. The numerical results are presented in Section 3.3.1.6.

3.3.1.1 Discretization of the adjoint of the continuous problem

To find the equations satisfied by the adjoint state p, let us see the control problem as a minimization problem under the constraint $i\hbar\dfrac{\partial\psi}{\partial t} = H\,\psi$ and $\psi(t=0) = \psi_0$. We emphasize that this is only a formal method to determine the adjoint problem and to compute the gradient. We shall skip the rigorous verification that the adjoint problem we find is indeed the correct one and that it yields the correct gradient. Using definitions given by Equation (3.4), we will write in this section the Hamiltonian H in the form : $H = H_{rot} + H_{laser}$. We recall that only H_{laser} depends on $\mathcal{E}$. Let us first introduce some definitions and notations that we will use throughout

this section. For $\mathcal{E}_1, \mathcal{E}_2 \in V_t = L^2([0,T],\mathbb{R})$ we define the scalar product

$$\langle \mathcal{E}_1 | \mathcal{E}_2 \rangle_{t,C} = \int_0^T \mathcal{E}_1(t)\mathcal{E}_2(t)dt,$$

and for $\phi_1, \phi_2 \in V_\theta = L^2([0,2\pi],\mathbb{C})$ the scalar product

$$\langle \phi_1 | \phi_2 \rangle_{\theta,C} = \int_0^\pi \Re\left(\phi_1(\theta)\overline{\phi_2(\theta)}\right)\sin\theta d\theta.$$

We also define for $\psi_1, \psi_2 \in V = L^2([0,T]\times[0,2\pi],\mathbb{C})$ the scalar product

$$\langle \psi_1 | \psi_2 \rangle_{t,\theta,C} = \int_0^T\int_0^\pi \Re\left(\psi_1(\theta,t)\overline{\psi_2(\theta,t)}\right)\sin\theta d\theta dt.$$

The subscript C aims at recalling the "continuous" nature of the scalar product, in comparison with the semi-discrete or the discrete ones which will be used later on. We emphasize that when differentiating functions with complex variables we consider these functions as two-variable functions and more precisely the complex variable is taken as an element of $\mathbb{R}^2$. For a given laser field $\mathcal{E}$, we denote by $\psi_\mathcal{E}$ the solution of Equation (3.1). Therefore, we define $\widetilde{J}$ using the criterion J as : $J(\mathcal{E}) = \widetilde{J}(\psi_\mathcal{E})$. Thus for $\mathcal{E} \in V_t$ and $(\psi, p) \in V^2$ we write the Lagrangian $\mathcal{L}^C$ of the continuous problem as follows :

$$\begin{aligned}\mathcal{L}^C(\mathcal{E},\psi,p) &= \widetilde{J}(\psi) + \left\langle \left(i\hbar\frac{\partial}{\partial t} - H_{rot} - H_{laser}\right)\psi \middle| p \right\rangle_{t,\theta,C} \\ &\quad + \left\langle \psi(\cdot,t=0) - \psi^0 | p(\cdot,t=0)\right\rangle_{\theta,C}. \end{aligned} \tag{3.11}$$

With standard, but tedious, calculations mainly based upon the linearity of the scalar product and that of the operators H_{rot} and H_{laser}, with an integration by part and with

$$\widetilde{J}'(\psi)\cdot\delta\psi = \frac{1}{T}\int_0^T\int_0^\pi \Re\left[2\overline{\psi}\cos\theta\delta\psi\right]\sin\theta d\theta dt, \tag{3.12}$$

we obtain $\dfrac{\partial\mathcal{L}^C}{\partial\psi}(\mathcal{E},\psi_\mathcal{E},p)\cdot\delta\psi$, which when set to zero gives the adjoint problem

$$\begin{cases} i\hbar\dfrac{\partial p}{\partial t} = H_{rot}p + H_{laser}p - \dfrac{2}{T}\psi_\mathcal{E}\cos\theta, \\ p(T) = 0. \end{cases} \tag{3.13}$$

We next formally compute the gradient $\nabla^C J$ using the Lagrangian $\mathcal{L}^C$. When using $\psi = \psi_\mathcal{E}$ the expression of the Lagrangian is $\mathcal{L}^C(\mathcal{E},\psi_\mathcal{E},p) = \widetilde{J}(\psi_\mathcal{E}) = J(\mathcal{E})$, thus we get

$$J'(\mathcal{E})\cdot\delta\mathcal{E} = \frac{\partial\mathcal{L}^C}{\partial\psi}(\mathcal{E},\psi_\mathcal{E},p)\cdot\frac{\partial\psi_\mathcal{E}}{\partial\mathcal{E}}\cdot\delta\mathcal{E} + \frac{\partial\mathcal{L}^C}{\partial\mathcal{E}}(\mathcal{E},\psi_\mathcal{E},p)\cdot\delta\mathcal{E},$$

which is simplified into $J'(\mathcal{E}) \cdot \delta\mathcal{E} = \dfrac{\partial \mathcal{L}^C}{\partial \mathcal{E}}(\mathcal{E}, \psi_\mathcal{E}, p) \cdot \delta\mathcal{E}$ when p is the adjoint state $p_\mathcal{E}$. Therfore, the gradient $\nabla^C J = \dfrac{dJ}{d\mathcal{E}}$ is obtained by

$$\begin{aligned}
\left\langle \nabla^C J | \delta\mathcal{E} \right\rangle_{t,C} &= \frac{\partial \mathcal{L}^C}{\partial \mathcal{E}}(\mathcal{E}, \psi_\mathcal{E}, p_\mathcal{E}) \cdot \delta\mathcal{E} \\
&= \left\langle -\left(\frac{\partial H_{laser}}{\partial \mathcal{E}}(\mathcal{E}) \cdot \delta\mathcal{E} \right) \psi_\mathcal{E} \middle| p_\mathcal{E} \right\rangle_{t,\theta,C} \\
&= \int_0^T \int_0^\pi \Re\left[\left(\mu_0 \cos\theta + \mathcal{E}\left[\alpha_\parallel \cos^2\theta + \alpha_\perp \sin^2\theta \right] \right) \psi_\mathcal{E} \overline{p_\mathcal{E}} \delta\mathcal{E} \right] \sin\theta d\theta dt.
\end{aligned}$$

The discretization of Equation (3.7) is done with an operator splitting method,

$$\left\{ \begin{array}{l} \psi^0, \\ \psi^{n+1} = e^{-\frac{i}{\hbar}\frac{\Delta t}{2} H^n_{laser}} e^{-\frac{i}{\hbar}\Delta t H_{rot}} e^{-\frac{i}{\hbar}\frac{\Delta t}{2} H^n_{laser}} \psi^n, \end{array} \right. \tag{3.14}$$

where H^n_{laser} is the time-dependent operator taken at time step t^n. Using this scheme to discretize the linear part of Equation (3.13) we obtain

$$\left\{ \begin{array}{l} p^N = 0, \\ p^{n-1} = e^{\frac{i}{\hbar}\frac{\Delta t}{2} H^n_{laser}} e^{\frac{i}{\hbar}\Delta t H_{rot}} e^{\frac{i}{\hbar}\frac{\Delta t}{2} H^n_{laser}} p^n + \dfrac{2}{T} \psi^n_\mathcal{E} \cos\theta \dfrac{\Delta t}{i\hbar}. \end{array} \right. \tag{3.15}$$

In addition, we use the same schemes for the time and space discretizations as the ones used for computing $J(\mathcal{E})$. More precisely, we use for time discretization a simple Riemann rule integration scheme. For the integration in θ, the method used is the Simpson rule

$$\begin{aligned}
\int_0^\pi g(\theta) d\theta &= \frac{\pi}{2N} \sum_{k=0}^{2N} \alpha_k g(\theta_k) \\
&= \frac{\Delta\theta}{3} \left[g(\theta_0) + 4 \sum_{k=0}^{N-1} g(\theta_{2k+1}) + 2 \sum_{k=0}^{N-2} g(\theta_{2k+2}) + g(\theta_{2N}) \right],
\end{aligned} \tag{3.16}$$

where $\Delta\theta = \dfrac{\pi}{2N}$ and where $(\theta_k)_{k=0,2N}$ are the equally-spaced integration points. Therefore the discretization of the gradient (3.14) reads, with an approximation in $(\Delta t)^2$ and in $(\Delta\theta)^4$,
$\left\langle \nabla^C J | \delta\mathcal{E} \right\rangle_{t,C} =$

$$\sum_{n=0}^{N-1} \sum_{k=0}^{2M} \Re\left[\mu_0 \cos\theta_k + \mathcal{E}^n \left[\alpha_\parallel \cos^2\theta_k + \alpha_\perp \sin^2\theta_k \right] \psi^n_k \overline{p}^n_k \right] \delta\mathcal{E}^n \Delta t \alpha_k \sin\theta_k \Delta\theta. \tag{3.17}$$

3.3.1.2 Adjoint calculus on the semi-discretized equations

The discretization of the time-dependent Schrödinger Equation (3.7) is given by (3.14) while the discretization of the criterion (again by the Riemann rule integration scheme) yields the semi-discrete Lagrangian

$$\begin{aligned}\mathcal{L}^{SD}(\mathcal{E},\Psi,P) &= \frac{1}{T}\sum_{n=0}^{N-1}\int_0^{\pi}\|\psi^n\|_{\mathbb{C}}^2\cos\theta\sin\theta d\theta\Delta t + \left\langle\psi^0-\psi_0|p^0\right\rangle_{\theta,C} \\ &\quad + \left\langle \Psi^S - e^{-\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}}e^{-\frac{i}{\hbar}\Delta t H_{rot}}e^{-\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}}\Psi\middle| P\right\rangle_{t,\theta,SD}, \quad (3.18)\end{aligned}$$

where $\Psi^S=(\psi^1,\dots,\psi^N)$, $\Psi=(\psi^0,\dots,\psi^{N-1})$ and $P=(p^0,\dots,p^{N-1})$ are elements of $(V_\theta)^N$ and where $\mathcal{E}=(\mathcal{E}^0,\dots,\mathcal{E}^{N-1})$ is an element of $\mathbb{R}^N$. The scalar product $\langle\cdot|\cdot\rangle_{\theta,C}$ is the one given in the previous section and the scalar product $\langle\cdot|\cdot\rangle_{t,\theta,SD}$ is given by $\langle\Psi_1|\Psi_2\rangle_{t,\theta,SD}=\sum_{n=1}^{N}\langle\psi_1^n|\psi_2^n\rangle_{\theta,C}\,\Delta t$ with $\Psi_1,\Psi_2\in(V_\theta)^N$. We also define for $\mathcal{E}_1,\mathcal{E}_2\in\mathbb{R}^n$ the scalar product $\langle\mathcal{E}_1|\mathcal{E}_2\rangle_{t,SD}=\sum_{n=1}^{N}\mathcal{E}_1^n\mathcal{E}_2^n$. For a given laser field $\mathcal{E}$ we denote $\Psi_{\mathcal{E}}$ the solution of Equation (3.14). As in the previous section, by computing $\frac{\partial\mathcal{L}^{SD}}{\partial\Psi}(\mathcal{E},\Psi_{\mathcal{E}},P)\cdot\delta\Psi$ and then by setting it to zero we get the following discrete adjoint problem :

$$\begin{cases} p^{N-1}=0, \\ p^{n-1}=e^{\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}^n}e^{\frac{i}{\hbar}\Delta t H_{rot}}e^{\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}^n}p^n-\frac{2}{T}\psi_{\mathcal{E}}^n\cos\theta. \end{cases} \quad (3.19)$$

And the gradient is obtained by :
$\frac{\partial\mathcal{L}^{SD}}{\partial\mathcal{E}}(\mathcal{E},\Psi_{\mathcal{E}},P_{\mathcal{E}})\cdot\delta\mathcal{E}$

$$\begin{aligned}&= \sum_{n=0}^{N-1}\left\langle -\frac{i\Delta t}{\hbar}\frac{\partial H_{laser}^n}{\partial\mathcal{E}_n}\delta\mathcal{E}^n e^{-\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}^n}e^{-\frac{i}{\hbar}\Delta t H_{rot}}e^{-\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}^n}\psi^n\middle| p^n\right\rangle_{\theta,C} \\ &= \sum_{n=0}^{N-1}\left\langle -\frac{i\Delta t}{\hbar}\frac{\partial H_{laser}^n}{\partial\mathcal{E}_n}\delta\mathcal{E}^n e^{-\frac{i}{\hbar}\frac{\Delta t}{2}H_{laser}^n}\psi^{n+1}\middle| p^n\right\rangle_{\theta,C}. \quad (3.20)\end{aligned}$$

Thus, with an approximation at the order $(\Delta\theta)^4$, we obtain
$\left\langle\nabla^{SD}J|\delta\mathcal{E}\right\rangle_{t,SD}=$

$$-\sum_{n=0}^{N-1}\sum_{k=0}^{2M}\Re\left[\frac{i}{\hbar}\psi_k^{n+1}\overline{p}_k^n\left(\mu_0\cos\theta_k+\mathcal{E}^n\left[\alpha_{\|}\cos^2\theta_k+\alpha_{\perp}\sin^2\theta_k\right]\right)\delta\mathcal{E}^n\right](\Delta t)^2\sin\theta_k\alpha_k\Delta\theta. \quad (3.21)$$

3.3.1.3 Comparison of the continuous and the semi-discretized approaches

In order to understand which of the formulae (3.17) or (3.21) is more accurate, we give below some illustrative example. Although very basic, this example allows one to understand the fundamental difference between formulae (3.17) and (3.21). Let us argue on the following Schrödinger equation :

$$\begin{cases} i\hbar\dfrac{\partial\psi}{\partial t} = \psi\,\mathcal{E}(t)\cos\theta, \\ \psi(0) = \psi^0, \end{cases} \tag{3.22}$$

(obtained by simply setting H_0 to zero in (3.7)) with the criterion written in the form

$$J(\mathcal{E}) = \widetilde{J}(\psi) = \frac{1}{T}\int_0^T\int_0^\pi f(\psi)\sin\theta d\theta dt.$$

The first way to proceed is the one we have followed in Section 3.3.1.1, namely by discretizing the adjoint equation. For Equation (3.22), basic calculus shows that the adjoint equation is given by

$$\begin{cases} i\hbar\dfrac{\partial p}{\partial t} = \mathcal{E}(t)p\cos\theta - \overline{f'}(\psi_\mathcal{E}), \\ p(T) = 0, \end{cases} \tag{3.23}$$

which once discretized with the same scheme as the one used for the direct equation, yields

$$\begin{cases} p^N = 0, \\ p^{n-1} = e^{\frac{i}{\hbar}\frac{\Delta t}{2}H^n_{laser}}e^{\frac{i}{\hbar}\Delta t H_{rot}}e^{\frac{i}{\hbar}\frac{\Delta t}{2}H^n_{laser}}p^n - \overline{f'}(\psi^n_\mathcal{E})\dfrac{\Delta t}{i\hbar}. \end{cases} \tag{3.24}$$

We compute the gradient of the criterion,

$$\nabla J(\mathcal{E}) = \frac{1}{T}\int_0^T\int_0^\pi f'(\psi_\mathcal{E})\frac{\partial\psi}{\partial\mathcal{E}}\sin\theta d\theta dt,$$

where $\dfrac{\partial\psi}{\partial\mathcal{E}}$ solves

$$\begin{cases} i\hbar\dfrac{\partial}{\partial t}\left(\dfrac{\partial\psi}{\partial\mathcal{E}}\right) = \psi\cos\theta + \mathcal{E}(t)\cos\theta\dfrac{\partial\psi}{\partial\mathcal{E}}, \\ \left.\dfrac{\partial\psi}{\partial\mathcal{E}}\right|_{t=0} = 0. \end{cases} \tag{3.25}$$

Thus, by using Equation (3.23) and integration by part we obtain

$$\langle\nabla J|\delta\mathcal{E}\rangle_{t,C} = -\int_0^T\left(\int_0^\pi\psi\overline{p}\cos\theta\sin\theta d\theta\right)\delta\mathcal{E}dt. \tag{3.26}$$

We now discretize this integral by the Riemann scheme which yields

$$\int_0^T \left(\int_0^\pi \psi \overline{p} \cos\theta \sin\theta d\theta \right) \delta\mathcal{E} dt = \sum_{n=0}^{N-1} \left(\int_0^\pi \psi^n \overline{p}^n \cos\theta \sin\theta d\theta \right) \delta\mathcal{E}^n \Delta t,$$

and thus the following approximation of the gradient :

$$\langle \nabla J(\psi) | \delta\mathcal{E} \rangle_{t,C} = -\sum_{n=0}^{N-1} \left(\int_0^\pi \psi^n \overline{p}^n \cos\theta \sin\theta d\theta \right) \delta\mathcal{E}^n \Delta t, \tag{3.27}$$

is the exact analogous of formula (3.17). Using this Riemann discretization scheme, the numerical error is controlled by the following estimate :

$$\left| \int_0^T g(t)dt - \sum_{n=0}^{N-1} g^n \Delta t \right| \leq T\Delta t \|g'\|_{L^\infty}. \tag{3.28}$$

Applying this result to $g = \left(\int_0^\pi \psi \overline{p} \cos\theta \sin\theta d\theta \right) \delta\mathcal{E}$, we obtain the control of the numerical error of the approximation (3.27) of the gradient

$$\begin{aligned} |\varepsilon_{\Delta t}^C| &\leq T\Delta t \left\| \frac{\partial}{\partial t} \left[\left(\int_0^\pi \psi \overline{p} \cos\theta \sin\theta d\theta \right) \delta\mathcal{E} \right] \right\|_{L^\infty} \\ &\leq T\Delta t \left\| \left[\frac{\partial}{\partial t} \left(\int_0^\pi \psi \overline{p} \cos\theta \sin\theta d\theta \right) \right] \delta\mathcal{E} + \left(\int_0^\pi \psi p \cos\theta \sin\theta d\theta \right) \frac{\partial}{\partial t} (\delta\mathcal{E}) \right\|_{L^\infty} \\ &\leq CT\Delta t \left(\|\delta\mathcal{E}\|_{L^\infty} + \left\| \frac{\partial}{\partial t} (\delta\mathcal{E}) \right\|_{L^\infty} \right), \end{aligned} \tag{3.29}$$

where the constant C depends on norms of $\psi|_{t=0}$ and $\mathcal{E}$ but not on $\delta\mathcal{E}$.

On the other hand, if we now discretize the equation and the criterion, we obtain as in (3.20)

$$\langle \nabla J | \delta\mathcal{E} \rangle_{t,SD} = \sum_{n=0}^{N-1} \left(\int_0^\pi \psi^{n+1} p^n \cos\theta \sin\theta d\theta \right) \delta\mathcal{E}^n \Delta t. \tag{3.30}$$

Applying the same numerical analysis, we see that the error in the approximation of the gradient is now obtained by setting $g = \int_0^\pi f'(\psi)\delta\psi \sin\theta d\theta$ in (3.28), which yields

$$\left| \varepsilon_{\Delta t}^{SD} \right| \leq T\Delta t \left\| \int_0^\pi \left(\frac{\partial}{\partial t} \left[f'(\psi)\delta\psi \right] \right) \sin\theta d\theta \right\|_{L^\infty}. \tag{3.31}$$

Now

$$\frac{\partial}{\partial t} \left(f'(\psi)\delta\psi \right) = f''(\psi) \frac{\partial \psi}{\partial t} \delta\psi + f'(\psi) \frac{\partial (\delta\psi)}{\partial t},$$

where $\frac{\partial \psi}{\partial t} = \frac{1}{i\hbar}(\mathcal{E}\psi\cos\theta)$ and $\frac{\partial(\delta\psi)}{\partial t} = \frac{1}{i\hbar}((\delta\mathcal{E})\psi\cos\theta + \mathcal{E}x(\delta\psi))$. It follows that

$$\left\| \int_0^\pi \left(\frac{\partial}{\partial t}\left[f'(\psi)\delta\psi \right] \right) \sin\theta d\theta \right\|_{L^\infty} \leq C\|\delta\mathcal{E}\|_{L^\infty},$$

where C only depends on norms on $\psi|_{t=0}$ and $\mathcal{E}$. Therefore

$$\left|\varepsilon_{\Delta t}^{SD}\right| \leq CT\Delta t\|\delta\mathcal{E}\|_{L^\infty}. \tag{3.32}$$

Comparing this estimate to (3.29), we see that the control in (3.32) is better, in particular for variations $\delta\mathcal{E}$ of $\mathcal{E}$ that have large variations in time, which will precisely be the case for us (oscillatory laser fields). It is therefore expected that *in our case* the adjoint calculus on the discrete equation will yield a better accuracy for the computation of the gradient than the approach consisting in discrtizing the continuous adjoint equation. Let us emphasize that the main difference between the two approaches is the following formal (non rigorous) integration by parts :

$$\int_0^T f'(\psi)\delta\psi \approx \int_0^T \left(\frac{\partial p}{\partial t}\right)\delta\psi \approx \int_0^T p\left(\frac{\partial \delta\psi}{\partial t}\right) \approx \int_0^T p\psi\delta\mathcal{E},$$

which is done before or after discretization and thus allows one to have the control of the error basically either by

$$\frac{\partial}{\partial t}\left(f'(\psi)\delta\psi\right) \approx f'(\psi)\delta\mathcal{E} \tag{3.33}$$

or by

$$\frac{\partial}{\partial t}\left(p\psi\delta\mathcal{E}\right) \approx p\psi\frac{\partial}{\partial t}\left(\delta\mathcal{E}\right). \tag{3.34}$$

In the case (3.33) the numerical error of integration is reported on ψ and $\delta\psi$ while in the case (3.34) the numerical error of integration is directly reported on $\delta\mathcal{E}$. Figure 3.7 summarizes the main ideas presented here.

3.3.1.4 Adjoint calculus on the fully discretized equations

In this section we begin by discretizing Equation (3.7) both in time and in θ-space and then do the adjoint calculus. The numerical propagation of the θ operator H_θ and the laser operator H_{laser} can be written in the matrix form

$$\Psi^{n+1} = A_\theta^n B A_\theta^n \Psi^n,$$

where Ψ^n is the vector (ψ_k^n), where A_θ^n is the diagonal matrix of the laser operator propagation, and where B is the matrix corresponding to the θ operator propagation.

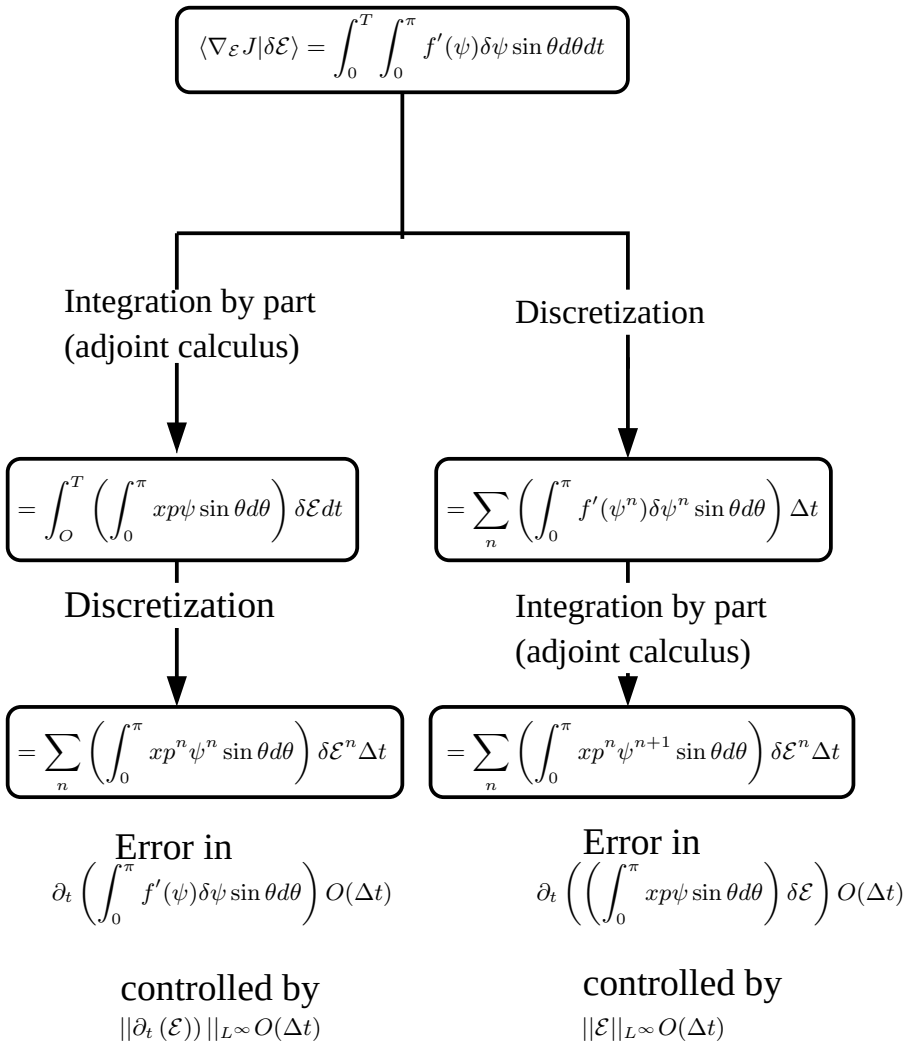


FIG. 3.7 – Comparaison of the two approaches of Section 3.3.1.1 and 3.3.1.2 to compute the gradient.

Only the matrix A_θ^n depends on the laser field $\mathcal{E}$.
We write the discrete Lagrangian as follows :

$$\begin{aligned}\mathcal{L}^D(\mathcal{E},\underline{\underline{\Psi}},\underline{\underline{P}}) &= \sum_{n=0}^{N-1}\sum_{k=0}^{2M}\frac{1}{T}\|\psi_k^n\|_{\mathbb{C}}^2\cos\theta\sin\theta\alpha_k\Delta\theta\Delta t \\ &\quad + \left\langle \underline{\underline{\Psi^S}} - \underline{\underline{\Psi^M}}|\underline{\underline{P}}\right\rangle_{t,D} + \left\langle \Psi^0 - \Psi_0|P^0\right\rangle_{\theta,D}, \end{aligned} \tag{3.35}$$

where

$$\underline{\underline{\Psi^S}} = (\Psi^1,\dots,\Psi^N),$$
$$\underline{\underline{\Psi^M}} = (A_\theta^0 B A_\theta^0\Psi^0,\dots,A_\theta^{N-1}BA_\theta^{N-1}\Psi^{N-1}),$$

and

$$\underline{\underline{P}} = (P^0,\dots,P^{N-1})$$

are elements of $\left(\mathbb{R}^{2M}\right)^N$ and where $\mathcal{E} = (\mathcal{E}^0,\dots,\mathcal{E}^{N-1})$ is an element of $\mathbb{R}^N$.
The scalar product $\langle\cdot|\cdot\rangle_{t,D}$ is given for $\underline{\underline{\Psi_1}},\underline{\underline{\Psi_2}} \in \left(\mathbb{R}^{2M}\right)^N$ by

$$\left\langle \underline{\underline{\Psi_1}}\middle|\underline{\underline{\Psi_2}}\right\rangle_{t,D} = \sum_{n=1}^{N}\langle\Psi_1^n|\Psi_2^n\rangle_{\theta,D}\,\Delta t$$

with $\langle\Psi_1|\Psi_2\rangle_{\theta,D} = \sum_{n=0}^{2M}\Re\left(\psi_{1k}\overline{\psi_{2k}}\right)\alpha_k\sin\theta\Delta\theta$ when $\Psi_1,\Psi_2\in\mathbb{R}^{2M}$.
Therefore, equation $\dfrac{\partial\mathcal{L}^D}{\partial\underline{\underline{\Psi}}}(\mathcal{E},\underline{\underline{\Psi_\mathcal{E}}},\underline{\underline{P}})\cdot\delta\underline{\underline{\Psi}} = 0$, we get

$$\begin{cases} P^{N-1} = 0 \\ p_k^{n-1} = \left[\left(\overline{A_\theta^n}\,\overline{B}\,\overline{A_\theta^n}\right)_k^T P^n\right]_k - \alpha_k\frac{2}{T}\psi_k^n\cos\theta. \end{cases} \tag{3.36}$$

The gradient is obtained by
$\dfrac{\partial\mathcal{L}^D}{\partial\mathcal{E}}(\mathcal{E},\underline{\underline{\Psi_\mathcal{E}}},\underline{\underline{P_\mathcal{E}}})\cdot\delta\mathcal{E}$

$$\begin{aligned} &= -\left\langle\left(\frac{\partial\underline{\underline{\Psi^M}}}{\partial\mathcal{E}}\right)\cdot\delta\mathcal{E}|\underline{\underline{P}}\right\rangle_{t,D} \\ &= -\sum_{n=0}^{N-1}\sum_{k=0}^{2M}\Re\left(\left[\left(\frac{\partial A_\theta^n}{\partial\mathcal{E}^n}BA_\theta^n + A_\theta^nB\frac{\partial A_\theta^n}{\partial\mathcal{E}^n}\right)\Psi^n\right]_k\overline{p}_k^n\delta\mathcal{E}^n\right)\alpha_k\sin\theta\Delta\theta\Delta t, \end{aligned}$$

where $\dfrac{\partial A_\theta^n}{\partial\mathcal{E}^n}$ is the matrix obtained by differentiating the matrix A_θ^n. We obtain for the gradient formula

$$\left\langle \nabla^D J | \delta\mathcal{E} \right\rangle_{t,D} =$$

$$\sum_{n=0}^{N-1} \sum_{k=0}^{2M} \Re\left[\left(\mu_0 \cos\theta_k + \mathcal{E}^n \left[\alpha_\parallel \cos^2\theta_k + \alpha_\perp \sin^2\theta_k \right] \right) \widetilde{\psi_k^n} \overline{p}_k^n \delta\mathcal{E}^n \right] \alpha_k \sin\theta \Delta\theta \Delta t,$$

where $\widetilde{\Psi^n} = \left(\frac{\partial A_\theta^n}{\partial \mathcal{E}^n} B A_\theta^n + A_\theta^n F^{-1} A B F \frac{\partial A_\theta^n}{\partial \mathcal{E}^n} \right) \Psi^n$. This formula is to be compared with (3.17) and (3.21).

3.3.1.5 Computing the gradient using Automatic Differentiation tools

In this section we briefly present another method to compute the gradient, which uses the Automatic Differentiation tool O∂yssée [19, 37]. Automatic Differentiation tools can be seen as black boxes taking as input a program computing a cost function $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and giving as output another program computing the gradient $\frac{\partial f(x)}{\partial x}$.
O∂yssée is able to use two modes : the tangent mode and the adjoint mode, which is similar to the adjoint method. We emphasize that the cost of the gradient computation is proportional to n with the tangent mode (as with finite differences) and it is proportional to m with the adjoint mode. Thus, the tangent mode has to be used when $n \ll m$ and the adjoint mode has to be used when $n \gg m$. For our problem we have $m = 1$ and $1 \ll n \leq 70$, so we use only the adjoint mode. As in the direct program we have 50 000 iterations, the adjoint program needs a lot of memory to run. In order to reduce the size of memory needed, we have modified the adjoint program by deleting the temporary variables in the linear parts of the program. Table 3.2 gives an idea of the size of the direct code and that of the adjoint code. The calculation times refer to a Pentium II, 466 Mhz Celeron with 128 Mb RAM running with Linux.

Tab. 3.2 – Technical requirements with O∂yssée.

	direct code	standard adjoint code	post-processed adjoint code
Size (lines)	433	2075	1190
Memory needed	12 Ko	520 Mo	103 Mo
Time (CPU)	60 s	—	141 s

3.3.1.6 Numerical results

The purpose of this section is to compare numerically the different methods presented above for computing the gradient. For the numerical tests we use one laser field

of the form $\mathcal{E}(t) = E\sin(\omega t + \phi)$. The gradient with respect to the parameters E, ω and ϕ is denoted as

$$\nabla J = \begin{pmatrix} \nabla_E J \\ \nabla_\omega J \\ \nabla_\phi J \end{pmatrix}.$$

We have computed the gradient using the methods presented in the previous sections, more precisely the continuous approach (C), the semi-discrete approach (SD), the discrete approach (D), and O∂yssée (AD). We have also computed the gradient using the finite differences approach (FD), where for each variable x ($x = E, \omega, \phi$)

$$\nabla_x^{FD} J = \lim_{\delta x \to 0} \frac{J(x + \delta x) - J(x)}{\delta x}.$$

The gradient given by FD has been computed with different values of δx to make sure that we have reached the $\delta x \to 0$ limit. Next we compare the gradient obtained using the different approaches with the gradient obtained using the finite differences approach, which is therefore taken as a reference value. For each method we will compute the relative error

$$e_{E,\omega,\phi} = \left| \frac{\nabla_{E,\omega,\phi} J - \nabla^{FD}_{E,\omega,\phi} J}{\nabla^{FD}_{E,\omega,\phi} J} \right|.$$

The comparison is done for both low and high frequencies, using two different representative points
$(E, \omega, \phi) = (10^{11}\ \mathrm{W/cm^2}, 500\ \mathrm{cm^{-1}}, 0)$ and $(E, \omega, \phi) = (10^{11}\ \mathrm{W/cm^2}, 4000\ \mathrm{cm^{-1}}, 0)$, respectively.
Table 3.3 shows that all the methods we have presented in this section give good results compared to the finite differences method. We can also see on this table that the best results are obtained using O∂yssée, where we have a better precision than with the other methods. In general the precision is increased by at least one order. We can also see on this table that the results agree with the comparison we made in Section 3.3.1.3 between the continuous approach and the semi-discrete approach, except for the ∇_E component.

Let us now take the results given by automatic differentiation as a reference and make the same comparison with the other methods as we have done with the finite differences approach. On Table 3.4 we see that, compared to the AD approach, the best results are those given by the discrete approach (again except for the component ∇_E). We recall that with the discrete approach we make the adjoint calculus on the fully discretized equation and that the automatic differentiation tools make also adjoint calculus on the fully discretized equation with some implementation differences. We also emphasize that for the component ∇_E we obtain results which are different from the results we obtain with the components ∇_ω and ∇_ϕ. We still unable to explain such a difference.

In practice, the size of the parameter vector for our problem can go up to 70, so we can only use an adjoint based method and not the finite differences one. Indeed the

CPU time needed to compute the gradient depends on the parameter vector size for the finite differences method and is independent of this size for the other methods presented below. More precisely, the CPU time needed for these other methods is about 3 times the one needed to compute the criterion. For implementation, the continuous approach and the semi-discrete approach are easiest to implement than the discrete approach. Finally, for O∂yssée, let us recall that even if it gives automatically the gradient, some post-processing of the adjoint code is needed before running it.

TAB. 3.3 – Relative error, with respect to the FD, of the gradient.

(a) : gradient computed at $(E, \omega, \phi) = (10^{11}, 500, 0)$

	FD	AD	C	SD	D
e_E	0.	$25. \times 10^{-8}$	$89. \times 10^{-6}$	$20. \times 10^{-3}$	$41. \times 10^{-3}$
e_ω	0.	$92. \times 10^{-6}$	$13. \times 10^{-4}$	$47. \times 10^{-5}$	$32. \times 10^{-5}$
e_ϕ	0.	$18. \times 10^{-6}$	$81. \times 10^{-5}$	$25. \times 10^{-5}$	$25. \times 10^{-5}$

(b) : gradient computed at $(E, \omega, \phi) = (10^{11}, 4000, 0)$

	FD	AD	C	SD	D
e_E	0.	$75. \times 10^{-5}$	$68. \times 10^{-5}$	$20. \times 10^{-3}$	$40. \times 10^{-3}$
e_ω	0.	$11. \times 10^{-5}$	$76. \times 10^{-4}$	$51. \times 10^{-4}$	$25. \times 10^{-4}$
e_ϕ	0.	$22. \times 10^{-3}$	$41. \times 10^{-3}$	$33. \times 10^{-3}$	$26. \times 10^{-3}$

TAB. 3.4 – Relative error, with respect to the AD, of the gradient.

(a) : gradient computed at $(E, \omega, \phi) = (10^{11}, 500, 0)$

	FD	AD	C	SD	D
e_E	$25. \times 10^{-8}$	0.	$89. \times 10^{-6}$	$20. \times 10^{-3}$	$41. \times 10^{-3}$
e_ω	$92. \times 10^{-6}$	0.	$14. \times 10^{-4}$	$57. \times 10^{-5}$	$22. \times 10^{-5}$
e_ϕ	$18. \times 10^{-6}$	0.	$79. \times 10^{-5}$	$23. \times 10^{-5}$	$27. \times 10^{-5}$

(b) : gradient computed at $(E, \omega, \phi) = (10^{11}, 4000, 0)$

	FD	AD	C	SD	D
e_E	$75. \times 10^{-5}$	0.	$73. \times 10^{-6}$	$20. \times 10^{-3}$	$41. \times 10^{-3}$
e_ω	$11. \times 10^{-5}$	0.	$75. \times 10^{-4}$	$50. \times 10^{-4}$	$24. \times 10^{-4}$
e_ϕ	$22. \times 10^{-3}$	0.	$18. \times 10^{-3}$	$10. \times 10^{-3}$	$36. \times 10^{-4}$

3.3.2 Evolutionary Algorithms

This section presents the stochastic algorithms that we have used for the orientation problem which belong to the family of Evolutionary Algorithms (EAs). Their common feature is to imitate the principle of natural evolution. This section is organized as follows : Section 3.3.2.1 briefly introduces EAs and their basic terminology and gives also a short state of the art in EAs while Section 3.3.2.2 presents more precisely the EAs that we have implemented for the orientation problem.

3.3.2.1 Introduction to Evolutionary Algorithms

This section briefly explains the basic steps of an EA. The problem is to optimize a given *objective function* f over a given search space. A population of individuals (i.e., a P-uple of points in the search space) undergoes some artificial Darwinian evolution based on the *fitness* F of each individual. The fitness of an individual is directly related to the value of the objective function of this individual (a typical example of a fitness function is the objective function itself, denoted by J in our work). The evolution operators applied to the individuals are defined upon the so-called *genotype space* noted E. It may be different from the definition space of the fitness called the *phenotype space*. The choice of this genotype space is the *representation*.
Figure 3.8 illustrates the framework of an EA : after an initialization of the population (generally a uniform random initialization) the fitness of each individual is computed. This is the *evaluation step*. Then, the loop of the algorithm called *a generation* is made up of the following steps :

- *Stopping criterion* : a basic stopping criterion is when the maximum number of generations fixed by the user is reached.
- *Selection* : the selection operator selects among the parents those who will generate offsprings, the genitors. There exists several selection operators, either of deterministic or of stochastic type. All of them are based on the fitness of the individuals and implement the first phase of Artificial Darwinism : the fittest allowed to reproduce.
- *Creation of new individuals* : there are basically two ways to create new individuals in the population from the genitors, namely the *crossover* and the *mutation*. These variation operators are stochastic operators : the crossover is a stochastic operator from E^k into E (typically $k = 2$), it is a recombination of k parents, and the mutation operator is a stochastic operator from E into E.
- *Evaluation* : for each offspring the fitness is computed.
- *Replacement* : this operator discriminate among the individuals of the current population those who will be the parents for the next generation. This operator, like the selection operator, is based on the fitness of the individuals and implements the second step of Darwin's theory : survival of the fittest.

Despite the common features of all EAs, several trends can be discriminated, mainly due to historical differences. We will only detail here the instances of EAs we have been using, refering to [3] and references therein for a complete description. The four main branches are (in alphabetical order) :

- The evolutionary programming (EP), originally developed in California to evolve finite state machines.
- The evolution strategies (ESs) developed in Germany to solve numerical optimization problems for real search spaces. The genotype space is the phenotype space, namely a subset of $\mathbb{R}^N$. A precise description is given below.
- The genetic algorithms (GAs) developed in Michigan to study some adaptation mechanisms of populations for biology. These algorithms have later been used for optimization problems. More precision are given below.

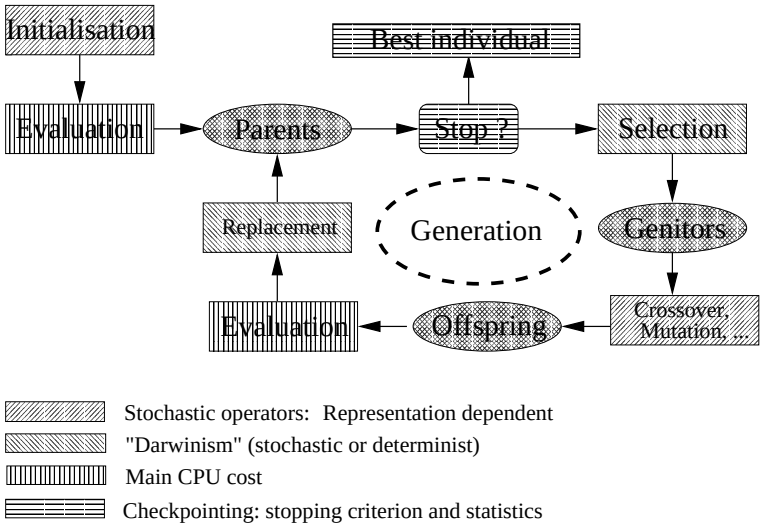


FIG. 3.8 – General EA scheme.

– The genetic programming (GP), which has appeared more recently, consists in evolving tree structures.

The Canonical GA :

The genotype space is $\{0,1\}^n$, the selection operator is the so-called roulette wheel, where the probability P_{X_p} to select the individual X_p is proportional to the fitness $F(X_p)$:

$$P_{Xp} = \frac{F(X_p)}{\sum_{i \in Population} F(X_i)}.$$

The crossover operator replaces some bits in the first parent string by the corresponding bits from the second parent, and the mutation operator randomly flips a bit of the parent. The replacement is *generational* : the offsprings at the generation n become the parents of the generation $n+1$. Modern GAs are commonly used with any kind of representation as long as crossover and mutation can be defined.

The ES :

The ES [31] have been designed to optimize real functions, thus the natural search space is $\mathbb{R}^N$. The individuals undergo Gaussian mutations, namely addition of zero-mean Gaussian variables of standard deviation σ. The particularity of ES is that the parameter σ is a part of the genetic information. For a so-called isotropic ES, an individual is of the form $I = (x_1, \ldots, x_N, \sigma)$ and, for a non isotropic ES,

$I = (x_1, \ldots, x_N, \sigma_1, \ldots, \sigma_N)$ (there also exists a third type of ES not discussed here, the correlated ES). Consequently the mutation parameters are subjected to recombination and mutation as well. More precisely, the *adaptive mutation* takes place in two steps, first a mutation of the mutation parameters, second a mutation of object variables x_i. For an isotropic ES the two steps are

$$\begin{aligned} \sigma^{(t+1)} &= \sigma^{(t)} \exp(\tau_0 N(0,1)), \\ x_i^{(t+1)} &= x_i^{(t)} + N_i(0, \sigma^{(t+1)}), \end{aligned}$$

and, for a non isotropic ES,

$$\begin{aligned} \sigma_i^{(t+1)} &= \sigma_i^{(t)} \exp(\tau_0 N(0,1) + \tau N_i(0,1)), \\ x_i^{(t+1)} &= x_i^{(t)} + N_i(0, \sigma_i^{(t+1)}), \end{aligned}$$

where $N(0,1)$ stands for a Gaussian random variable. The crossover operator selects randomly two parents, $(x_1^1, \ldots, x_N^1, \sigma_1^1, \ldots, \sigma_N^1)$ and $(x_1^2, \ldots, x_N^2, \sigma_1^2, \ldots, \sigma_N^2)$, to produce an offspring $(x_1^{q_1}, \ldots, x_N^{q_N}, \sigma_1^{q_1}, \ldots, \sigma_N^{q_N})$ where $q_i = 1$ or $q_i = 2$ with equal probability. This crossover operator can also involve all individuals in the population, this is a *global crossover*. The replacement operator is strictly deterministic, based on the rank. For example, if μ (respectively λ) is the number of parents (respectively offsprings), $(\mu, \lambda) - ES$ selects the parents for next generation by taking the μ best offsprings and $(\mu + \lambda) - ES$ selects the parents for next generation by taking the μ best among the λ offsprings and μ parents.
It is now commonly accepted that the incorporation of specific knowledge, of the problem to optimize, by means of representation and specific operators, is the best way and the only way to enhance the performances of an EA. But, when using an EA without introducing some specificities of the problem, ES is generally the most efficient EA for parametric optimization. ES, like GA, are implemented on the EOlib class library available from [18].

3.3.2.2 The algorithms used

The orientation problem is a minimization problem on a real space of size $7N$, where N is the number of laser fields to superpose. We use two kinds of EAs : the first one is based on a classical GA with a real representation (roulette wheel selection and *barycentric* or *multi-point* crossover [28]) and the second one is the ES described above and taken from EOlib [18].
The first algorithm is an improved GA, adding some specific operators and some specific features, which are known to improve the performances of GAs. We will name this algorithm in the sequel EGA (for Enhanced GA). *Niching* and *Rescaling* are two specific features of this algorithm. Rescaling is a way to avoid some bias in the roulette wheel selection ; niching is to avoid that all the population concentrates on a region of the search space (see [28] for more precisions). Then, the mutation strength on EGA decreases with the number of generations. A specific gradient mutation operator is also used (EGA-CG), replacing the parent by the result of a few iterations of a conjugated gradient algorithm using the parent as initial value. The

purpose of such an operator is to accelerate the convergence by taking advantages of a gradient algorithm.
We have tested EGA, EGA-CG and ES on several test functions taken from the literature (Sphere, Rosenbrock and Shekel functions). We refer to [36] for the details. We present here shortly some conclusions of these tests. First, for all the functions tested, a comparison with a classical GA has shown that EGA, EGA-CG, and ES converge more often, and faster than GA. Moreover, they are able to improve continuously their precision whereas GA stops at some non-zero distance of the solution. Second, the tests have confirmed that the gradient mutation operator accelerates the convergence, except for too "chaotic" functions. Third, the test cases have helped us for a crucial point of EAs, namely the setting of the parameters, which is specific to each function. Several trends can be discriminated for the setting, mainly taken from the literature and confirmed with test cases. As we have built our own EGA, it is difficult to give succinctly the parameters to set. With respect to ES, three important steps are given : First, the probability to mutate an individual is greater than the probability to cross two individuals (typically $p_{mut} = 0.8$ and $p_{cross} = 0.2$). Second, the size of the population is typically $(7, 49) - ES$ and the number of parents should be increased if the number of local minima increases. Third, the initial mutation strength σ should also be increased when the number of local minima increases.

3.4 Results for the orientation problem

A preliminary study of the orientation problem with the purely deterministic PRLS and BFGS algorithms (see Section 3.3.1), for the differentiable criterion (3.10), showed the need to use stochastic methods. Indeed, these algorithms converge after a few iterations towards a local minimum close to the initial guess : the cost function presents numerous local minima. We know from the literature and from test cases that for such functions, ES, EGA, and EGA-CG perform better. As far as EGA-CG is concerned, using it to minimize the criterion (3.10) does not improve the results in a significant way. More precisely, our best results have been obtained without the gradient mutation operator. However, using the gradient algorithm after EGA can improve the result of the optimization, as we will see in Section 3.4.1.
We present in this section the main results obtained with our algorithms on the orientation problem. The sequel of this section is organized as follows : in Section 3.4.1, we give the fields we have obtained by minimizing criteria (3.9) and (3.10). For both criteria we give the best results. Next, we explain how the addition of the CG at the end of the EGA improves the optimization of the criterion (3.10). In Section 3.4.2, we introduce a new hybrid criterion in order to approach both goals of Section 3.2.3 : obtaining at some given time a good orientation and keeping it as long as possible. Then, in Section 3.4.3, we present results obtained by a different form of laser fields. This form of laser field, named a *train of kicks* is a succession of fields of *kick* form presented in Section 3.4.1. As we will see, these fields really improve the results of Section 3.4.1 on both criteria (3.9) and (3.10).

3.4.1 Optimized fields for (3.9) and (3.10)

All the results presented in the sequel have been obtained by optimizing upon a superposition of two or three lasers in order to better understand the physical meaning of the results. Indeed, our trials for optimizing (3.10) on a superposition of ten laser fields have given results shown on Figure 3.10, which are not sufficiently easy to understand and interpret. We have therefore left this strategy aside.
Figures 3.11 and 3.12 show the optimized fields and their instantaneous criterion $j(t)$ obtained respectively with criteria (3.9) and (3.10). They have been obtained with a non-isotropic ES and with EGA, respectively. However, let us emphasize that the two algorithms give similar results. Indeed, EGA has given fields and instantaneous criterion of the same form as the ones shown on Figure 3.11 and ES has also given results of the form shown on Figure 3.12.

As it may be noticed on Figure 3.11, the minimum value of $j(t)$, namely -0.46, is less than that on Figure 3.12 but the orientation does not last as long, which is expected in view of the criterion chosen. The first instantaneous criterion (Figure 3.11) is what we call a *narrow* $j(t)$ (see Section 3.2.3) and the second one (Figure 3.12) is what we call a *wide* $j(t)$. As for the fields, the first field is what we call in Section 3.2.4 a *$(\omega, 2\omega)$* field and the second one is what we call a *kick* field. Table 3.5 shows the parameters of this latter field. The fact that a field of the form of a *kick* is a very efficient field for optimizing the criterion 3.10 is one of our most striking result from a physical viewpoint. It is reported and commented on in [17]. In the latter reference, the *$(\omega, 2\omega)$* field is also analyzed. As explained above,

Tab. 3.5 – Parameters of the optimized pulse with 3 laser fields.

n	$\mathcal{E}$ (W/cm^2)	ω (cm^{-1})	ϕ (π rad)	t_0 (ps)	t_1 (ps)	t_2 (ps)	t_3 (ps)
1	1.01364×10^{08}	1389.541	1.98066	0.	0.312024	0.613023	1.193727
2	2.99976×10^{12}	500.051	1.82249	0.075077	0.270294	0.838110	1.562814
3	2.99989×10^{12}	500.000	0.82337	0.109518	0.235767	0.808280	1.080066

using the EGA-CG does not improve the results. However, CG is useful for a local search and we have tested how it could improve the result when used only at the end of a stochastic search. For this purpose we have first made an optimization on criterion (3.10) using EGA (the result is presented on Figure 3.13 with dotted lines) and then, we have applied the BFGS algorithm (the gradient has been computed with O∂yssée) using the laser field so obtained as an initial guess for the conjugate gradient algorithm. After 100 CG iterations, the criterion is improved as may be seen on Figure 3.13 with solid lines. Such a result reconfirms that CG is useful for the local improvement search.

3.4.2 Results for the hybrid criterion

In view of the results of the previous section, it is a natural idea to introduce a new criterion aimed at approaching two goals together : obtaining at some given time a good orientation and keeping it as long as possible. Thus, we basically define a new

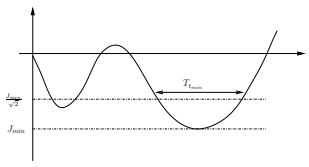


FIG. 3.9 – Construction of the hybrid criterion.

criterion

$$J = J_{min} - J_{kept} + |J_{min} + J_{kept}|, \tag{3.37}$$

where $J_{min} = \min_{t\in[0,T]} j(t)$ and $J_{kept} = \dfrac{T_{t_{min}}}{T}$ where $T_{t_{min}}$ is the length the connex component of $\{t \in [0,T] \mid J_{min} \leq j(t) \leq \frac{J_{min}}{\sqrt{2}}\}$ including $t_{min} = sup\{t \mid J(t) = J_{min}\}$ (see Figure 3.9). This criterion is a sum of three terms. The first one, J_{min}, measures the way the molecule is oriented. The second one, J_{kept}, measures how long the orientation is kept. The third part, $|J_{min} + J_{kept}|$, is a penalty term to ensure that J_{min} and $-J_{kept}$ are simultaneously minimized.
On Figure 3.14, we show a field obtained with this criterion and which is a succession in time of two fields with a short overlay time (see Section 3.2.4). For the physical meaning of such a result, we refer to [2].

3.4.3 Results for the train of kicks

An other idea consists of starting with a field previously classified as a kick shape and using a succession of such fields in order to orient the molecule. The purpose of the optimization is thus to find the good delay between two successive *kicks*. Indeed we hope that by kicking several times the molecule we can lower the instantaneous criterion. The results are quite interesting : Figure 3.15 (a), is the result of an optimization of the criterion (3.10) with ES and it clearly illustrates the idea of kicking several times the molecule. This result is also interesting because the instantaneous criterion remains for a long time under the value -0.2. Figure 3.15 (b) is the result of the optimization with the criterion (3.9). The criterion value (-0.82) is the best value we have ever had. However, the production of such fields remains an experimental challenge.

3.5 Conclusion and future directions

We have implemented and tested various strategies for the optimization of the laser field to be used for the orientation of the HCN molecule.

The best results have been obtained using evolutionary algorithms rather than purely deterministic algorithms such as gradient-like algorithms. However, in the case where the criterion is differentiable, we have shown that gradient like algorithms can efficiently complement the EA, not necessary when being used throughout the

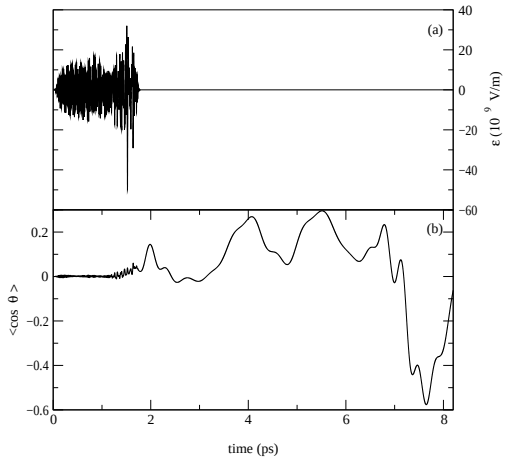


FIG. 3.10 – Results obtained by optimizing upon 10 laser fields. In this figure and the following ones (3.11 to 3.14), the electric field is shown on top while $< \cos\theta >$ which measures the instantaneous orientation of the molecule is shown on bottom. Time evolves from left to right at the same scale.

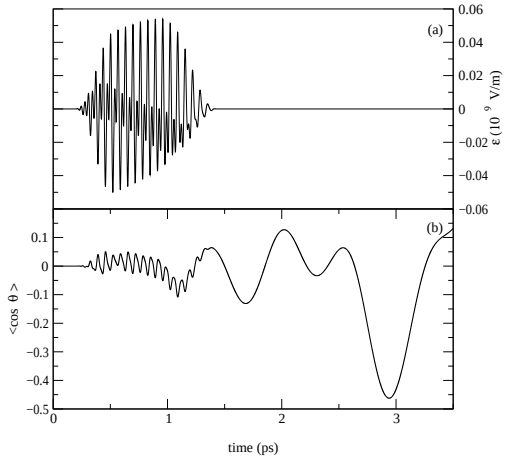


FIG. 3.11 – Best result for $J = \min_{t\in[0,T]} j(t)$.

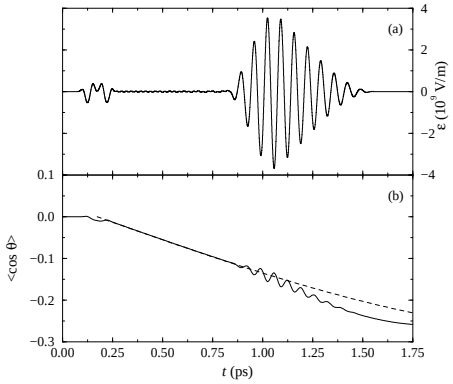


FIG. 3.12 – Best result for $J = \frac{1}{T}\int j(t)\,dt$.

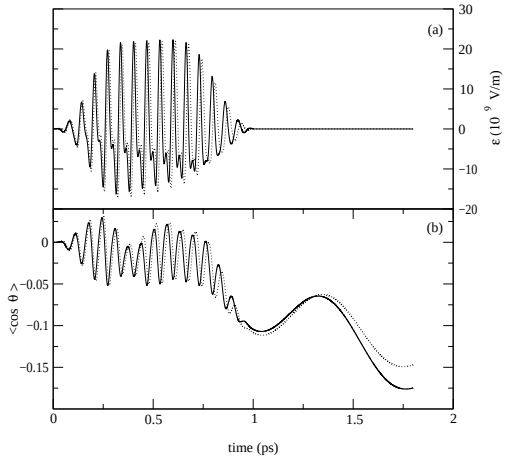


FIG. 3.13 – Optimization by CG after optimization by GA.

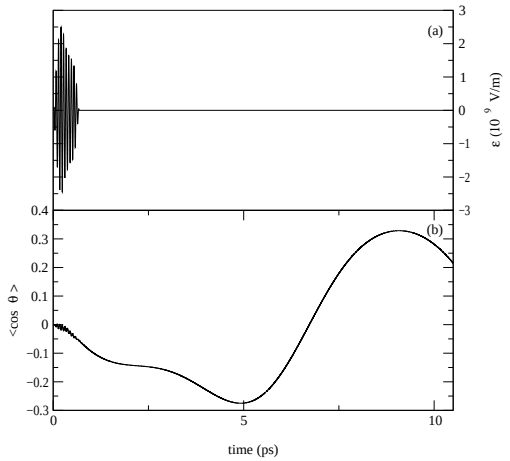


FIG. 3.14 – Best result for the hybrid criterion given by equation 3.37.

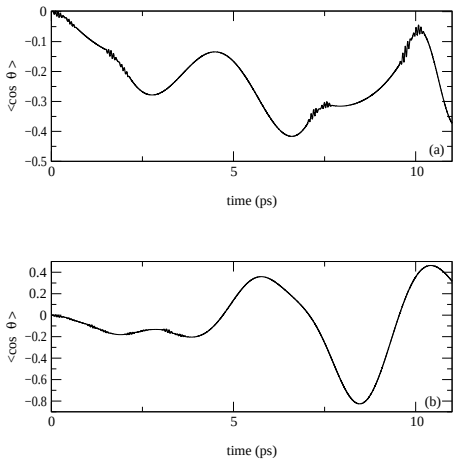


FIG. 3.15 – (a) : Best result for $J = \frac{1}{T}\int j(t)\,dt$ with the train of kicks. (b) : Best result for $J = \frac{1}{T}\int j(t)\,dt$ with the train of kicks.

generations as mutations operators (the genetic algorithms with mutation by gradient have not yielded a real benefit in our specific case), but when being used as a final step in the optimization, once the population has been optimized by EA.

In order to understand how to compute the gradient of the criterion when needed, we have performed many tests, together with a numerical analysis on a toy equation related to our case of interest. They both show that the most efficient strategy (amenable in any case) is to compute the gradient by adjoint calculus on the discretized form of the equation or, if one does not fear a tedious post-processing work, to compute the gradient with an automatic differentiation tool.

As far as the choice of the criterion is concerned, we have tested many criteria, depending upon our physical aims. A multicriteria approach has also been implemented.

From the physical standpoint, our results have allowed us to identify two specific forms of laser fields that are most promising for the future : the $(\omega, 2\omega)$ field [14, 26] and the *kick* field [15, 22]. Definite conclusions about the efficiently of these fields are yet to be obtained and will be the purpose of some of our work in the future. It is anyway to be emphasized that such physically relevant fields have been obtained through our optimization methodology used as a blind tool, *i.e.*, without any specification of this form of fields. This is sufficient to give us some hope and confidence both in the physical and in the mathematical validity of our methodology.

Acknowledgements

The financial support of the *Action Concertée Incitative Jeunes Chercheurs* from the French Ministry of Research and Technology is gratefully acknowledged. CLB would like to thank Herschel Rabitz and André Bandrauk for enlightening discussions. The expertise of Marc Schöenauer on evolutionary strategies has been intrumental. Finally, Gabriel Turinici and Yvon Maday are to be thanked for constant stimulating interactions.

References

[1] A. Assion, T. Baumer, M. Bergt, T. Brixnerand B. Kiefer, V. Seyfried, M. Strehle, and G. Gerber. Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses. *Science*, 282 :919–922, 1998.

[2] A. Auger, C. M. Dion, A. Ben Haj Yedder, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Numerical optimization of laser fields to control molecular orientation. submitted.

[3] Th. Bäck, D.B. Fogel, and Z. Michalewicz, editors. *Handbook of Evolutionary Computation*. Oxford University Press, Institute of Physics Publishing and Oxford University Press : Bristol and New York, 1997.

[4] J.M. Ball, J.E. Marsden, and M. Slemrod. Controllability for distributed bilinear systems. *SIAM J. Control Optim.*, 20 :575–597, 1982.

[5] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation Numérique : aspects théoriques et pratiques*. Springer, Berlin, 1997.

[6] C. Le Bris. Control theory applied to quantum chemistry : Some tracks. In *International Conference on Control of Systems Governed by PDEs (Nancy, March 1999)*, volume 8, pages 77–94. ESAIM PROC, 2000.

[7] C. Le Bris. Problématiques numériques pour la simulation moléculaire. In *Actes du 32ème Congrès national d'analyse numérique*, volume 9. ESAIM : Proceedings, September 2000.

[8] P. Brumer and M. Shapiro. Control of unimolecular reactions using coherent light. *Chem. Phys. Lett.*, 126 :541–546, 1986.

[9] P. Brumer and M. Shapiro. Laser control of chemical reactions. *Scientific American*, pages 34–39, March 1995.

[10] E. Cancès, C. Le Bris, and M. Pilot. Contrôle optimal bilinéaire sur une équation de Schrödinger. *Note aux Compte Rendu de l'Académie des Sciences*, pages 567–571, 2000.

[11] S. Chelkowski, M. Zamojski, and A. D. Bandrauk. Laser-phase directional control of photofragments in dissociative ionization of H_2^+ using two-color intense laser pulses. *Phys. Rev. A*, 63 :023409, 2001.

[12] C. E. Dateo and H. Metiu. Numerical solution of the time-dependent schrodinger equation in sperical coordinates by fourier-transformation methods. *J. Chem. Phys*, 95 :7392–7400, 1991.

[13] Académie des Sciences. Sciences aux temps ultracours ; de l'atoseconde aux petawatts. *Rapport sur la science et la technologie*, 2000, September.

[14] C. M. Dion, A. D. Bandrauk, O. Atabek, A. Keller, H. Umeda, and Y. Fujimura. Two-frequency ir laser orientation of polar molecules. numerical simulations for hcn. *Chem. Phys. Lett.*, 302 :215–223, 1999.

[15] C. M. Dion, A. Keller, and O. Atabek. Orienting molecules using half-cycle pulses. *Eur. Phys. J. D*, 14 :249–255, 2001.

[16] C. M. Dion, A. Keller, O. Atabek, and A. D. Bandrauk. Laser-induced alignment dynamics of HCN : Roles of the permanent dipole moment and the polarizability. *Phys. Rev. A*, 59 :1382–1391, 1999.

[17] C. M. Dion, A. Ben Haj Yedder, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Optimal laser control of orientation : The kicked molecule. submitted.

[18] EO. C++ class library, http ://eodev.sourceforge.net/.

[19] C. Faure and Y. Papegay. O∂yssée User's Guide Version 1.7. *Rapport Technique INRIA RT-0224*, 1998.

[20] M.D. Feit, J.A. Fleck, and A. Steiger. Solution of a Schrodinger equation by a spectral method. *J. Comput. Phys.*, 47 :412–433, 1982.

[21] B. Freidrich and D. R. Herschbach. On the possibility of orienting rotationally cooled polar molecules in an electric field. *Z. Phys. D*, 18 :153–161, 1991.

[22] N. E. Henriksen. Molecular alignment and orientation in short pulse laser fields. *Chem. Phys. Lett.*, 312 :196–202, 1999.

[23] K. Hoki and Y. Fujimura. Quantum control of alignment and orientation of molecules by optimized laser pulses. *Chem. Phys.*, 267 :187–193, 2001.

[24] R. S. Judson, K. K. Lehmann, H. Rabitz, and W. S. Warren. Optimal design of external fields for controlling molecular motion : Application to rotation. 223 :425–456, 1990.

[25] R. S. Judson and H. Rabitz. Teaching lasers to control molecules. *Phys. Rev. Lett.*, 68 :1500–1503, 1992.

[26] T. Kanai and H. Sakai. Numerical simulation of molecular orientation using strong, nonresonant, two-color laser fields. *J. Chem. Phys.*, 115 :5492–5497, 2001.

[27] J. J. Larsen, I. Wendt-Larsen, and H. Stapelfeldt. Controlling the branching ratio of photodissociation using aligned molecules. *Phys. Rev. Lett.*, 83 :1123–1126, 1999.

[28] Z. Michalewicz. *Genetic algorithms + data structure = evolution programs.* Springer, 1999.

[29] R. Numico, A. Keller, and O. Atabek. Laser-induced molecular alignment in dissociation dynamics. *Phys. Rev. A*, 52 :1298–1309, 1995.

[30] H. Sakai, C. P. Safvan, J. J. Larsen, K. M. Hilligsoe, K. Hald, and H. Stapelfeldt. Controlling the alignment of neutral molecules by a strong laser field. *J. Chem. Phys.*, 110 :10235–10238, 1999.

[31] H. P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[32] P. Tournois, D. Hulin, F. Amiranoff, G. Mourou, and D. Kaplan. Les implusions lasers ultra-brèves. *Compte rendu des* 3^{es} *Entretiens de la physiques*, 1998.

[33] G. Turinici. Controlabilité exacte de la population des états propres dans les systèmes quantiques bilinéaires. *Note aux Compte Rendu de l'Académie des Sciences*, pages 327–332, 2000.

[34] G. Turinici and H. Rabitz. Quantum wave function controllability. *Chem. Phys.*, 267 :1–9, 2001.

[35] G. Turinici and H. Rabitz. Wavefunction controllability in quantum systems. *Preprint*, 2001.

[36] A. Ben Haj Yedder. PhD thesis, Ecole Nationale des Ponts et Chaussées, in preparation.

[37] A. Ben Haj Yedder, E. Cancès, and C. Le Bris. Optimal laser control of chemical reactions using automatic differentiation. In George Corliss, Christèle Faure, Andreas Griewank, Laurent Hascoët, and Uwe Naumann (eds.), editors, *Proceedings of Automatic Differentiation 2000 : From Simulation to Optimization*, pages 203–213, New York, 2001. Springer-Verlag.

[38] W. Zhu and H. Rabitz. A rapid monotonically convergent iteration algorithm for quantum optimal control over the expectation value of a positive definite operator. *J. Chem. Phys.*, 109 :385–391, 1998.

Chapitre 4

Optimal Laser Control of Orientation : The Kicked Molecule

Ce chapitre est la reproduction d'un article paru dans *Physical Review A* [P2]. L'article présente l'un des premiers résultats trouvés et qui permis de retrouver un mécanisme d'orientation déjà connu et appelé *mécanisme de kick*. Ce résultat offre une nouvelle possibilité pour la réalisation de ce mécanisme et qui peut être plus facile à mettre en oeuvre expérimentalement.

Optimal Laser Control of Orientation : The Kicked Molecule

C. M. Dion[1,2] A. Ben Haj-Yedder[2] E. Cancès[2] A. Keller[1] C. Le Bris[2] O. Atabek[1]

[1] *Laboratoire de Photophysique Moléculaire du CNRS, Bâtiment 213, Campus d'Orsay, 91405 Orsay, France*

[2] *CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée, France*

Abstract: Using an optimal control scheme, based on genetic algorithms, to tailor a laser pulse, we find that molecular orientation can be achieved during and after the radiative interaction. The mechanism, which appears to be one of the possible ways leading to orientation, is based upon a kick imparted to the molecule by a sudden (with respect to molecular rotational motion), asymmetric laser pulse. We show how such pulses resulting from optimization can actually be produced experimentally and how the laser control of orientation could further be improved.

4.1 Introduction

Molecular orientation is not only a major concern in chemical reaction dynamics as an efficient cross-section enhancement device [1–5], but it is also determinant in controlling surface processing [6] or catalysis [7] and for nanoscale design by laser focusing of molecular beams [6, 8]. Basically, symmetry-breaking scenarios are referred to in achieving orientation : DC electric fields [4, 9], properly tailored microwave pulses [10], picosecond two-color phase-locked laser excitations [11], intense linearly-polarized IR pulses combining a fundamental frequency and its second harmonic resonant with a vibrational transition [12] or half-cycle pulses [13], both acting on polar molecules, eventually combining permanent-dipole- and polarizability-field interactions. Two-color schemes to control photofragment orientation have also been suggested [14, 15]. Among theoretical models that have so far been addressed, coherent and optimal laser control schemes have recently been proposed [10, 16, 17] as possible tools for orientation. This article is concerned with such an optimal laser control of molecular orientation using various genetic algorithms [18] to optimally tailor a pulse profile by properly summing up a number of individual fields characterized by their frequency, intensity, phase and temporal shape.

TAB. 4.1 – Molecular parameters.

Molecule	B (a.u.)	μ_0 (a.u.)	$\alpha_\parallel$ (a.u.)	$\alpha_\perp$ (a.u.)	T_{rot} (ps)
HCN[a]	6.6376×10^{-6}	1.1413	20.055	8.638	11.45
LiF[b]	5.9173×10^{-6}	2.5933	9.061	9.218	12.84

[a] From Ref. [19].

[b] Obtained from a quantum chemistry calculation [20].

4.2 Model

The HCN molecule, in its ground electronic state, taken as a rigid rotor, offers an illustrative example, in a model describing both permanent dipole (μ_0) and polarizability ($\alpha_\parallel$ and $\alpha_\perp$, parallel and perpendicular components) interactions. The complete molecule-plus-field Hamiltonian, at that level of approximation, is

$$\hat{H}(t) = B\hat{J}^2 - \mu_0 \mathcal{E}(t)\cos\theta - \left(\Delta\alpha\cos^2\theta + \alpha_\perp\right)\frac{\mathcal{E}^2(t)}{2}, \tag{4.1}$$

$\hat{J}^2$ being the angular momentum operator, B the rotational constant, θ the polar angle which defines orientation of the molecule with respect to the linearly-polarized electric field vector $\vec{\mathcal{E}}(t)$, at time t, and $\Delta\alpha \equiv \alpha_\parallel - \alpha_\perp$ the polarizability anisotropy. All relevant parameters are displayed in Table 4.1. The time evolution, starting from an isotropic initial distribution ($J = M_J = 0$, M_J being the projection of the total angular momentum J on the field polarization axis), is governed by the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\psi(\theta,\varphi;t) = \hat{H}(t)\psi(\theta,\varphi;t), \tag{4.2}$$

with φ the azimuthal angle. Equation 4.2 is solved using a third order split-operator method [21] in conjunction with a scheme developed by Dateo and Metiu [22, 23] for the angular variables. The measure of the orientation, taken as the evaluation function for the genetic algorithm, is the expectation value of $\cos\theta$,

$$\langle\cos\theta\rangle(t) = \int_0^\pi \cos\theta\sin\theta d\theta \int_0^{2\pi} |\psi(\theta,\varphi;t)|^2 d\varphi. \tag{4.3}$$

A sum of N individual linearly-polarized pulses,

$$\mathcal{E}(t) = \sum_{n=1}^{N} \mathcal{E}_n(t)\sin\left(\omega_n t + \phi_n\right), \tag{4.4}$$

builds up the electromagnetic field through the optimization procedure. The envelope functions $\mathcal{E}_n(t)$ are given sine-square forms,

$$\mathcal{E}_n(t) = \begin{cases} 0 & \text{if} \quad t \leq t_{n0} \\ \mathcal{E}_{n0} \sin^2 \left[\frac{\pi}{2} \left(\frac{t - t_{n0}}{t_{n1} - t_{n0}} \right) \right] & \text{if} \quad t_{n0} \leq t \leq t_{n1} \\ \mathcal{E}_{n0} & \text{if} \quad t_{n1} \leq t \leq t_{n2} \\ \mathcal{E}_{n0} \sin^2 \left[\frac{\pi}{2} \left(\frac{t_{n3} - t}{t_{n3} - t_{n2}} \right) \right] & \text{if} \quad t_{n2} \leq t \leq t_{n3} \\ 0 & \text{if} \quad t \geq t_{n3} \end{cases} \tag{4.5}$$

each pulse being characterized by a set of 7 adjustable parameters, namely its frequency ω_n, relative phase ϕ_n, maximum field amplitude $\mathcal{E}_{n0}$, together with 4 times determining its shape (origin t_{n0}, rise time $t_{n1} - t_{n0}$, plateau $t_{n2} - t_{n1}$, and extinction time $t_{n3} - t_{n2}$). How to tailor an electromagnetic field to reach the best possible orientation amounts to a parameter optimization problem involving $7 \times N$ variables aiming at the maximization of a target. An optimization criterion being specified, we are using a stochastic method by implementing a genetic algorithm [24] based on a floating point coding. For this genetic algorithm the mutation and crossover operators act upon the variables separately to preserve the physical meaning of the different parameters. In order to both explore the search space at the beginning and accelerate the convergence at the end, the mutation operator is adaptively modified by making its amplitude decrease with the iterations. During the optimization procedure, the parameters are constrained such that $\mathcal{I} = \epsilon_0 c \mathcal{E}^2/2 \leq 3 \times 10^{12}$ W/cm^2, $500 \leq \omega_n \leq 4000$ cm^{-1}, $t_{n0} < t_{n1} < t_{n2} < t_{n3}$, and $t_{n3} \leq 1.7$ ps. The latter constraint is meaningful in the sense that short-pulse *alignment* of molecules has already been demonstrated theoretically and experimentally [25–30]. Several criteria are possible for defining the target. The ideal behavior of $|\langle \cos\theta \rangle (t)|$ (nearly) equals to 1 for (almost) all time t, obviously being an unreachable physical challenge, some compromise have clearly to be accepted between orientation efficiency and/or duration. But before entering this delicate problem of the choice of criteria, one has to be guided (and encouraged) by some orientation dynamics resulting from the optimization of some intuitive elementary criterion.

From our previous experience [12] and to the best of our knowledge, orientation during the pulse has not yet been obtained, although the rotational distribution at the end of the pulse is crucial for any possible future orientation. This suggests taking as a starting point for our investigation the simple criterion

$$j \equiv \langle \cos\theta \rangle (t_f), \tag{4.6}$$

to be minimized, namely, the best possible orientation at the final interaction time $t_f = \sup_n (t_{n3})$. Even when restricting the investigation to this unique criterion, the optimization control scheme presents some severe limitations basically related with the large number of parameters to be sampled through the genetic algorithm by a rather time-consuming calculation of an evaluation function and the difficulty to retain clear physical interpretations from bulk results. This also serves as a preliminary study of the use of different genetic algorithms to tackle the problem of molecular orientation. Indeed, the

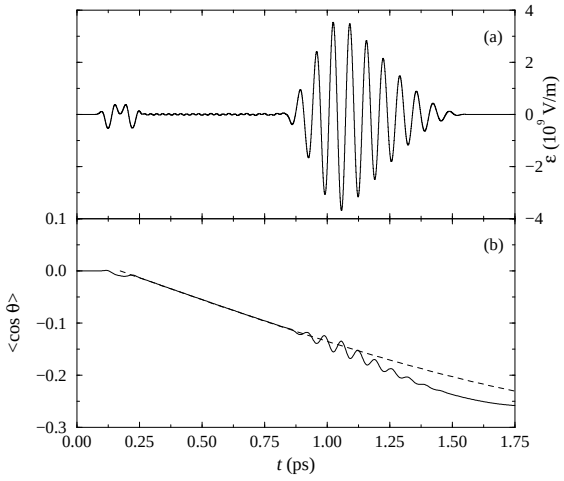


FIG. 4.1 – (a) Laser field derived from the optimal control calculation of the orientation for the HCN rigid rotor (see Table 4.2 for parameters). (b) Orientation expectation value $\langle \cos\theta \rangle$ with this field (solid line) and from the sudden-impact model [Eqs. (4.8) and (4.9)] with $\mathcal{A} = -0.465$, $\mathcal{B} = 1.18 \times 10^{-2}$, $\mathcal{C} = 8.92 \times 10^{-3}$, and $t_k = 0.172$ ps (dashed line).

difficulty of searching in a large parameter space (CPU-time consuming) requires the use of more advanced algorithms such as self-adaptive algorithms (evolution strategies) or genetic algorithms using gradient-mutation operators [31, 32].

4.3 Results

4.3.1 Optimization results

After a number of unsuccessful trials, an enlightening interpretation emerges from a calculation involving only $N = 3$ pulses : enlightening by the mechanism it suggests for one possible way of molecular orientation, but also by its simplicity that leads to a discussion of its experimental feasibility. Figure 4.1 shows the optimal laser field and the resulting orientation dynamics, while Tab. 4.2 collects the corresponding pulse parameters. Three periods can be retained when analyzing $\langle \cos\theta \rangle (t)$.

(i) An initial sudden (of approximately 0.25 ps duration, *i.e.*, much shorter than the rotational period of 11 ps) and asymmetric pulse imparts a kick to the molecule that induces the dynamics of orientation in a way very similar to the one in consideration when referring to half-cycle pulses [13]. The molecular response time is still not short enough for a noticeable quasi-instantaneous orientation to be observed within this period extending up to ~ 0.25 ps.

(ii) The orientation continues to develop during the second period ($\sim$ 0.25 ps to $\sim$ 1 ps) where all individual fields reach their plateau value ; two of them with equal

TAB. 4.2 – Parameters of the optimized laser pulse.

n	$\mathcal{I}_n$ (W/cm^2)	ω_n (cm^{-1})	ϕ_n (π rad)	t_{n0} (ps)	t_{n1} (ps)	t_{n2} (ps)	t_{n3} (ps)
1	1.01364×10^8	1389.541	1.98066	0.	0.312024	0.613023	1.193727
2	2.99976×10^{12}	500.051	1.82249	0.075077	0.270294	0.838110	1.562814
3	2.99989×10^{12}	500.000	0.82337	0.109518	0.235767	0.808280	1.080066

field amplitudes (within 4 digits accuracy) corresponding to rather high intensities of $\sim 3 \times 10^{12}$ W/cm^2, the third being 4 orders of magnitude smaller. But the most striking observation is that the strong field pulses present equal frequencies (within 4 digits accuracy) associated with a relative phase shift of π (within 3 digits accuracy), resulting in a quasi-absolute destructive interference that preserves the kick mechanism initiated in the first period.

(iii) A dozen field oscillations within a smoother envelope before complete switch-off at $t \approx 1.5$ ps characterizes the third period. Although the molecule is solicited back and forth it continues to be oriented by the effect of the initial kick $|\langle \cos\theta(t) \rangle|$ increasing with some additional wiggles which follow the field oscillations at ~ 500 cm^{-1} frequency.

Two points are to be emphasized for a better appreciation of these findings. First is that the unavoidable limitation of the parameter sampling space to $t_{n3} \leq 1.7$ ps, convenient for numerical calculations but already sudden with respect to the rotational period, can bias the optimization technique to force the kick mechanism. This actually is not the case since the optimized field lasts a time shorter than 1.7 ps and, even more convincing, the part of the field responsible for the kick is acting only over a period of ~ 0.25 ps. Second is that the interpretation of the kick mechanism is reinforced by the application of an impulsive "sudden-impact" model (as presented in Ref. [13]). This model basically reflects the fact that during the radiative interaction the molecular rotational motion can approximately be neglected. At lowest order, for a field taken from t_{ki} to t_{kf}, the fulfillment of the inequality

$$t_{kf} - t_{ki} \ll \frac{\hbar}{B\hat{J}^2} \tag{4.7}$$

(*i.e.*, short kick duration $t_{kf} - t_{ki}$ with respect to the rotational period) leads to the approximate solution of the time dependent Schrödinger equation

$$\begin{aligned} \psi(\theta, \varphi; t > t_k) \;\; = \;\; & \exp\left(-\frac{i}{\hbar} B\hat{J}^2 t\right) \exp\left[i \left(\mathcal{A}\cos\theta \right.\right. \\ & \left.\left. + \mathcal{B}\cos^2\theta + \mathcal{C}\right)\right] \psi(\theta, \varphi; t = t_{ki}), \end{aligned} \tag{4.8}$$

with $t_k = (t_{kf} - t_{ki})/2$, and where the coefficients $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are obtained from the integrated electric field as

$$\mathcal{A} = \frac{\mu_0}{\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}(t) dt, \tag{4.9a}$$

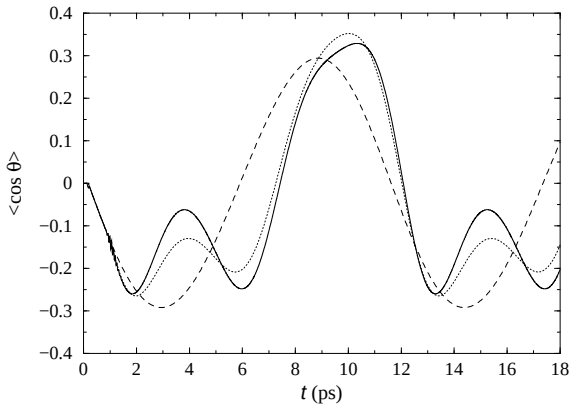


FIG. 4.2 – Same as Fig. 4.1(b), with an additional curve obtained from the sudden-impact model with a first kick with $\mathcal{A} = -0.465$, $\mathcal{B} = 1.18\times10^{-2}$, and $\mathcal{C} = 8.92\times10^{-3}$ at $t_k = 0.172$ ps and a second kick with $\mathcal{A} = -7.93 \times 10^{-2}$, $\mathcal{B} = 1.40$, and $\mathcal{C} = 1.06$ at $t_k = 1.059$ ps (dotted line).

$$\mathcal{B} = \frac{\Delta\alpha}{2\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}^2(t)dt, \tag{4.9b}$$

$$\mathcal{C} = \frac{\alpha_\perp}{2\hbar} \int_{t_{ki}}^{t_{kf}} \mathcal{E}^2(t)dt. \tag{4.9c}$$

Freezing rotational dynamics during the kick, by use of Eq. (4.8), leads to angular distributions within fairly good accuracy, as displayed in Fig. 4.1(b). The conclusion is that the optimized field of Fig. 4.1(a) definitely offers one way to produce orientation through a kick mechanism.

4.3.2 Field-free behavior

The long-time rotational dynamics of HCN radiated by the field of Fig. 4.1(a) is displayed in Fig. 4.2. The average of $\cos\theta$ continues on decreasing after the pulse is over during an additional time of 0.3 ps. The degree of orientation is improved up to $\langle\cos\theta\rangle \sim -0.26$ at $t \sim 1.85$ ps. Later on, the molecule remains oriented, to a lesser extent, up to 7.5 ps, when its direction with respect to the (now extinguished) laser polarization vector is back-reversed. The rotational population reached at the end of the pulse evolves by phase accumulation in such a way that, in this back direction (corresponding to positive values of $\cos\theta$), strikingly a much better orientation is obtained. Namely, $\langle\cos\theta\rangle$ reaches a maximum of 0.33 at $t \sim 10.3$ ps, and even more interesting is that the duration over which $\langle\cos\theta\rangle$ remains larger than 0.2 lasts for more than 3 ps. After a time ($t \approx 13$ ps) corresponding to the laser excitation duration plus a full rotational period has elapsed, the orientation is again in the forth direction with $\langle\cos\theta\rangle \sim -0.26$, as expected. It is to be noted that in a field-free situation, the rotational populations being only affected through

their evolving phases, the above-discussed patterns of $\langle \cos\theta \rangle$ will recurrently occur with the molecular rotational periodicity (until relaxation by spontaneous emission).

Two versions of the sudden-impact approximation are also illustrated in Fig. 4.2, in order to check the long-time consequences of the kick mechanism. The sudden-impact model as applied at $t_k = 0.172$ ps describes a single kick and turns out to be a rather modest approximation especially around $t = 4$ ps and 13 ps, when orientation is almost lost. The dynamics is much better described by referring twice to the sudden-impact model ; namely, at $t_k = 0.172$ ps and $t_k = 1.059$ ps. This shows that a double-kick mechanism prevails. The back and forth oscillations of the field in Fig. 4.1(a) between 0.85 ps and 1.50 ps, although intuitively hardly interpretable as a second kick, act as a sudden (*i.e.*, non-adiabatic with respect to the rotational period) excitation of the molecule producing an enhancement of the rotational populations. In particular, the $J = 2$ level populated up to 25% is now responsible of a sub-periodicity of $T_{\rm rot}/3$ in $\langle \cos\theta \rangle (t)$ resulting in the observed lost of orientation at 4 ps and 13 ps.

4.3.3 Temperature effects

Some previous studies in the literature have shown a rather fast decrease of the degree of alignment or orientation with increasing temperature [33–35]. The orientation is actually erased when Boltzmann averaging the rotational distributions due to the fact that a pulse optimized for the $J = 0$ $M = 0$ state is no more appropriate for orienting rotationally excited initial states. The average to be performed is given by

$$\begin{aligned} \langle\langle \cos\theta \rangle\rangle (t) &= Q^{-1} \sum_J \exp\left[\frac{-BJ(J+1)}{k_B T}\right] \\ &\quad \times \sum_{M=-J}^{J} \langle \cos\theta \rangle_{J,M} (t), \end{aligned} \tag{4.10}$$

where Q is the partition function

$$Q = \sum_J (2J+1) \exp\left[\frac{-BJ(J+1)}{k_B T}\right], \tag{4.11}$$

k_B and T being the Boltzmann constant and the rotational temperature, respectively. $\langle \cos\theta \rangle_{J,M} (t)$ describes the orientation dynamics for a given single initial state J, M calculated as previously through Eq. (4.3) by an exact wave packet propagation performed using Eq. (4.2) with $|J, M\rangle$ as an initial rotational state. The results for HCN under the effect of the excitation of the pulse of Fig. 4.1(a) (optimized for the $J = M = 0$ state alone) are shown, for three low rotational temperatures $T = 0$ K, 2K, and 5K, in Fig. 4.3. As expected, and in conformity with previous observations [34], the rapid increase of rotational population with temperature is at the origin of the decrease of the performances of the optimization technique based on the unique initial rotationless ground state.

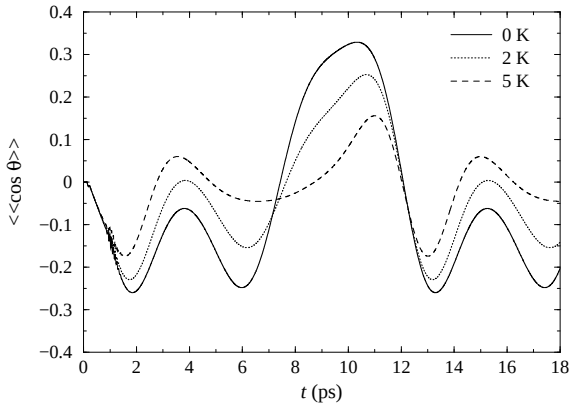


FIG. 4.3 – Thermal averages of the orientation expectation value $\langle \cos\theta \rangle$ [see Eq. (4.10)] for HCN submitted to the field given in Fig. 4.1(a).

4.3.4 The kick mechanism

If the kick mechanism is robust, the conclusions drawn for a given (molecule-plus-field) system would be successfully extended to a different one. First is the consideration of a different molecule with the following arrangements and expectations :

- larger permanent dipole moment, higher radiative coupling [see Eq. (4.9)] resulting in better orientation ;
- lighter molecule, higher rotational constant, less rotational levels populated under the effect of temperature resulting in a better resistance of the orientation to temperature effects.

The molecule which has been retained for this study is LiF, for which the relevant parameters characterizing the dipole moment as obtained from a quantum chemistry calculation [20] are given in Table 4.1. Although slightly heavier than HCN, this system presents the advantage, with respect to the kick mechanism, of a higher permanent dipole moment (more than twice the one of HCN). We also observe that the polarizability anisotropy $\Delta\alpha$ is much lower than the one of HCN. But, due to the fact that $\Delta\alpha$ is associated with a term in $\cos^2\theta$ in the laser-molecule coupling, Eq. (4.1), it is not (at least directly) responsible for orientation effects which are rather dominated by the permanent dipole interaction. The results, for the three previously considered temperatures, are gathered in Fig. 4.4. Two remarks are in order :

(i) a similar behavior to the case of HCN can definitely be considered as the robustness of the kick mechanism ;

(ii) but most encouraging is the obtainment, as expected, of a high degree of orientation, with a maximum for $\langle \cos\theta \rangle$ (at $T = 0$ K) close to 0.6. Even at 2 K, a degree of orientation better than $\langle\langle \cos\theta \rangle\rangle = 0.4$ is achieved during a time interval of about 1 ps.

Second is the consideration of different laser sources. Integrating all the information we get from the optimization procedure, together with some additional features in relation with the experimental feasibility of the laser pulse, we build up the field displayed

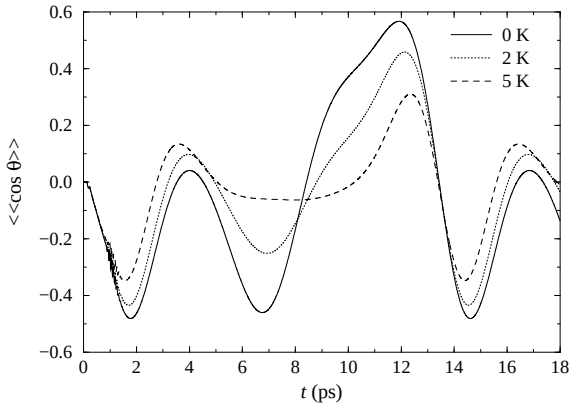


FIG. 4.4 – Same as Fig. 4.3, but for the LiF molecule.

TAB. 4.3 – Parameters of the constructed laser pulses.

n	$\mathcal{I}_n$	ω_n	ϕ_n	t_{n0}	t_{n1}	t_{n2}	t_{n3}
	W/cm^2	cm^{-1}	π rad	ps	ps	ps	ps
1	1×10^{13}	500	0	0.	0.20014	1.80125	2.00139
2	1×10^{13}	500	1	0.03336	0.23350	1.83461	2.03475

in Fig. 4.5(a) (parameters given in Table 4.3). More precisely, this is done by retaining the two major contributions to Eq. (4.5) as they result from the previous calculation for producing the first part lasting over ~ 0.25 ps and responsible for the initial kick. The radiative interaction is then switched off up to 1.75 ps to give enough time for the molecular response resulting in efficient orientation. A field identical to the one that produces the kick, but with a relative phase shift of π, is switched on at $t = 1.75$ ps. The orientation is attenuated after this second pulse, but such a double-pulse laser is experimentally reachable by an appropriate duplication and recombination of a single field of the general shape given by Eq. (4.5) and presenting a zero total time-integrated electric field amplitude. Once duplicated, one of the resulting pulses is time-delayed such that a phase shift of π be accurately achieved before recombination. The plateau behavior being washed out by the destructive interference, the resulting pulse is precisely the one of Fig. 4.5(a). The corresponding orientation dynamics is displayed in Fig. 4.5(b), together with the result of the sudden impact approximation which behaves rather closely and therefore advocates again for a kick mechanism. We emphasize that this result not only shows how to produce, in a different and presumably easier way than a half-cycle pulse, a realistic electromagnetic field imparting a kick to the molecule responsible for its orientation, but is also an additional signature of robustness. More precisely, both the robustness of the kick model itself (changes in the number, structure, and time delay of individual pulses, as long as the resulting field is "kick shaped") and numerical robustness with respect to small changes

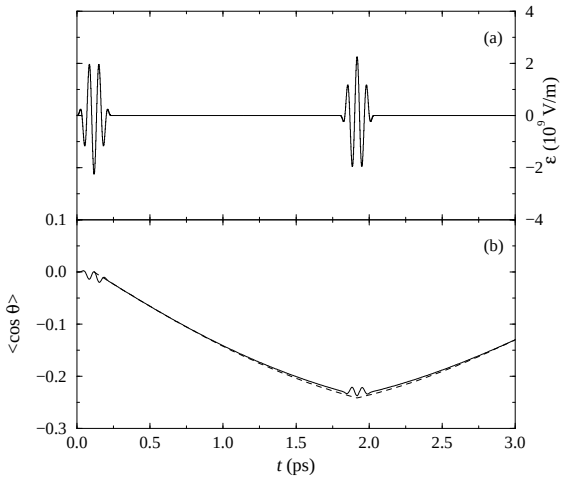


FIG. 4.5 – (a) Constructed laser field (see text and Table 4.3 for parameters). (b) Orientation expectation value $\langle \cos\theta \rangle$ with this field (solid line) and from the sudden-impact model with $\mathcal{A} = -0.483$, $\mathcal{B} = 0.226$, $\mathcal{C} = 0.171$, and $t_k = 0.117$ ps for first kick and $\mathcal{A} = 0.483$, $\mathcal{B} = 0.226$, $\mathcal{C} = 0.171$, and $t_k = 1.918$ ps for second kick (dashed line).

TAB. 4.4 – Parameters of the "optimal" constructed laser pulses.

n	$\mathcal{I}_n$	ω_n	ϕ_n	t_{n0}	t_{n1}	t_{n2}	t_{n3}
	W/cm^2	cm^{-1}	π rad	ps	ps	ps	ps
1	1×10^{13}	500	0	0.	0.20014	5.20360	5.40374
2	1×10^{13}	500	1	0.03336	0.23350	5.23696	5.43710

in the parameters (frequencies and intensities) have thus been successfully checked. This is to be added to the robustness with respect to the molecular system considered.

The long term behavior of this orientation dynamics of HCN is plotted in Fig. 4.6, together with the result of the sudden approximation which is, as expected, very close to the exact one. As previously suggested, the degree of orientation can finally be better optimized through the additional free parameter, namely the delay between the two kicks. This gives rise, for the field, to parameters which are given in Table 4.4, and for a resulting orientation dynamics also shown in Fig. 4.6. A remarkably good orientation is achieved for the rotationless ground state of HCN, namely $|\langle \cos\theta \rangle|$ is larger than 0.4 for time intervals over 2.5 ps.

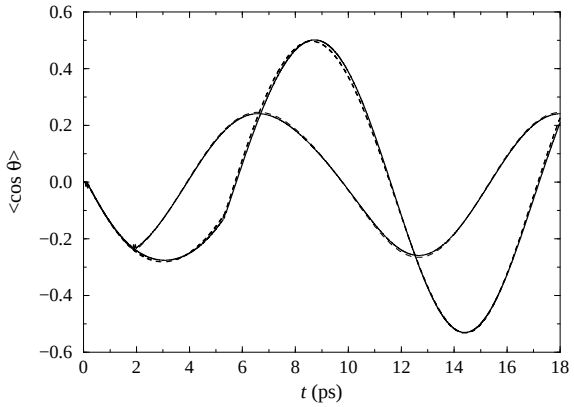


FIG. 4.6 – Orientation expectation value $\langle \cos\theta \rangle$ with the constructed "double kick" laser pulses. Thin lines : see Table 4.3 for parameters (solid line) and sudden-impact model with $\mathcal{A} = -0.483$, $\mathcal{B} = 0.226$, $\mathcal{C} = 0.171$, and $t_k = 0.117$ ps for first kick and $\mathcal{A} = 0.483$, $\mathcal{B} = 0.226$, $\mathcal{C} = 0.171$, and $t_k = 1.918$ ps for second kick (dashed line). Thick lines : same as thin lines, but with parameters from Table 4.4 and second kick of the sudden-impact model at $t_k = 5.330$ ps.

4.4 Conclusion

In conclusion, an optimal control scheme aiming at tailoring a laser pulse shape to achieve molecular orientation leads to a sudden asymmetric field that acts on the molecule like a unidirectional kick, a mechanism already referred to using half-cycle pulses. Although very accurately evidenced, such a mechanism is clearly not unique, other optimization criteria combining orientation efficiency and duration may presumably lead to efficient dynamics. It offers, however, apart from its physically sound and convincing interpretation, the advantage of robustness and experimental feasibility. Moreover, it may help as a guide to more advanced control scenarios. Namely, it suggests the use of a train of kicks acting in the same direction and progressively enhancing the orientation effect of the first, the time delay between them serving as a control parameter that basically takes into account the molecular response time to the laser excitation and the field-free evolution of the angular distribution. Work in this direction, undertaken together with consideration of more sophisticated optimization targets, is in progress in our group [32].

Acknowledgments: We acknowledge the financial support of the *Action Concertée Incitative Jeunes Chercheurs* from the French Ministry of Research and computing time allowed on a NEC SX5 by the *Institut du Développement et des Ressources Informatiques Scientifiques* of the CNRS.

References

[1] P. R. Brooks, Science **193**, 11 (1976).

[2] D. H. Parker, H. Jalink, and S. Stolte, J. Phys. Chem. **91**, 5247 (1987).

[3] A. H. Zewail, J. Chem. Soc. Faraday Trans. 2 **85**, 1221 (1989).

[4] H. J. Loesch and A. Remscheid, J. Chem. Phys. **93**, 4779 (1990).

[5] F. J. Aoiz, B. Friedrich, V. J. Herrero, V. S. Rábanos, and J. E. Verdasco, Chem. Phys. Lett. **289**, 132 (1998).

[6] T. Seideman, Phys. Rev. A **56**, R17 (1997).

[7] M. G. Tenner, E. W. Kuipers, A. W. Kleyn, and S. Stolte, J. Chem. Phys. **94**, 5197 (1991).

[8] H. Stapelfeldt, H. Sakai, E. Constant, and P. B. Corkum, Phys. Rev. Lett. **79**, 2787 (1997).

[9] B. Friedrich and D. R. Herschbach, Z. Phys. D **18**, 153 (1991).

[10] R. S. Judson, K. K. Lehmann, H. Rabitz, and W. S. Warren, J. Mol. Spectrosc. **223**, 425 (1990).

[11] M. J. J. Vrakking and S. Stolte, Chem. Phys. Lett. **271**, 209 (1997).

[12] C. M. Dion, A. D. Bandrauk, O. Atabek, A. Keller, H. Umeda, and Y. Fujimura, Chem. Phys. Lett. **302**, 215 (1999a).

[13] C. M. Dion, A. Keller, and O. Atabek, Eur. Phys. J. D **14**, 249 (2001).

[14] E. Charron, A. Giusti-Suzor, and F. H. Mies, Phys. Rev. A **49**, R641 (1994).

[15] M. Machholm and N. E. Henriksen, J. Chem. Phys. **111**, 3051 (1999).

[16] J. Ortigoso, Phys. Rev. A **57**, 4592 (1998).

[17] K. Hoki and Y. Fujimura, Chem. Phys. **267**, 187 (2001).

[18] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs* (Springer, Berlin, 1996), 3rd ed.

[19] C. Dion, Ph.D. thesis, Université Paris-Sud and Université de Sherbrooke (1999).

[20] *Gaussian 98* (Revision A.9), M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci,

C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle and J. A. Pople, Gaussian, Inc., Pittsburgh PA, 1998.

[21] M. D. Feit, J. A. Fleck, Jr., and A. Steiger, J. Comput. Phys. **47**, 412 (1982).

[22] C. E. Dateo and H. Metiu, J. Chem. Phys. **95**, 7392 (1991).

[23] R. Numico, A. Keller, and O. Atabek, Phys. Rev. A **52**, 1298 (1995).

[24] M. Vose, in *Foundations of Genetic Algorithms II*, edited by D. Whitley (Morgan Kaufmann, San Mateo, 1993).

[25] C. M. Dion, A. Keller, O. Atabek, and A. D. Bandrauk, Phys. Rev. A **59**, 1382 (1999b).

[26] N. E. Henriksen, Chem. Phys. Lett. **312**, 196 (1999).

[27] T. Seideman, J. Chem. Phys. **103**, 7887 (1995).

[28] J. H. Posthumus, J. Plumridge, M. K. Thomas, K. Codling, L. J. Frasinski, A. J. Langley, and P. F. Taday, J. Phys. B : At. Mol. Opt. Phys. **31**, L553 (1998a).

[29] J. H. Posthumus, J. Plumridge, L. J. Frasinski, K. Codling, A. J. Langley, and P. F. Taday, J. Phys. B : At. Mol. Opt. Phys. **31**, L985 (1998b).

[30] C. Ellert and P. B. Corkum, Phys. Rev. A **59**, R3170 (1999).

[31] A. Auger, A. Ben Haj-Yedder, E. Cancès, C. Le Bris, C. M. Dion, A. Keller, and O. Atabek, Math. Models Methods Appl. Sci. (in press).

[32] A. Ben Haj-Yedder, A. Auger, C. M. Dion, E. Cancès, A. Keller, O. Atabek, and C. Le Bris (in preparation).

[33] J. Ortigoso, M. Rodríguez, M. Gupta, and B. Friedrich, J. Chem. Phys. **110**, 3870 (1999).

[34] M. Machholm and N. E. Henriksen, Phys. Rev. Lett. **87**, 193001 (2001).

[35] M. Machholm, J. Chem. Phys. **115**, 10724 (2001).

Chapitre 5

Numerical optimization of laser fields to control molecular orientation

Ce chapitre est la reproduction d'un article à paraître dans *Physical Review A* [P4]. Dans cet article on présente les critères mesurant l'orientation de la molécule et utilisés par les méthodes d'optimisation. Le but de ces critères est de traduire l'objectif physique d'obtenir une orientation efficace et/ou assez longue. En particulier on présente des critères pouvant prendre en compte l'effet de la température. En effet, l'orientation obtenue en optimisation les autres critères se perd sous l'effet de la température.

Numerical optimization of laser fields to control molecular orientation

A. Ben Haj-Yedder[2] A. Auger[2] C. M. Dion[2] E. Cancès[2] A. Keller[1] C. Le Bris[2]
O. Atabek[1]

[1] *Laboratoire de Photophysique Moléculaire du CNRS, Bâtiment 213, Campus d'Orsay, 91405 Orsay, France*
[2] *CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée, France*

Abstract: A thorough numerical illustration of an optimal control scenario dealing with the laser-induced orientation of a diatomic molecule (LiF) is presented. Special emphasis is put on the definition of the various targets dealing with different orientation characteristics, identified in terms of maximum efficiency (i.e., molecular axis direction closest to the direction of the laser polarization vector), of maximum duration (i.e., the time interval during which this orientation is maintained), or of a compromise between efficiency and duration. Excellent post-pulse orientation is achieved by sudden, intense pulses. Thermal effects are also studied with an extension of the control scenarios to Boltzmann averaged orientation dynamics at $T = 5$ K.

5.1 Introduction

The relative orientation of the collision partners basically remains one of the most important features determining the outcome of their encounter. Experimental techniques based on hexapole state selection [1, 2] and the so-called "brute force" orientation in DC electric fields [3, 4] have already been developed to enable control over molecular orientation. Alignment and orientation of molecules are challenging issues covering a wide range of applications, extending from chemical reactivity enhancement and photofragment analysis with the possibility of separating the products, to surface processing, catalysis, or nanoscale design [5–10].

One of the basic mechanisms for laser-induced alignment (i.e., molecular axis parallel to the field polarization vector) can be understood in terms of the pendular states accommodated by the molecule-plus-field effective potential [11–13]. The laser control of alignment can be reached by an adiabatic transport of an initial isotropic rotational state to some pendular state trapping the molecule in well-aligned geometries. After the laser

is turned off, alignment can no longer be observed in this case. Alignment can also be obtained in field-free situations provided that the laser excitation be sudden with respect to the molecular rotational period, resulting in highly-excited rotational states [14]. Roughly speaking, alignment processes are rather well understood and experimentally documented [15–20].

Laser control of orientation (i.e., molecular axis in the same direction as the field polarization vector) requires in addition symmetry breaking mechanisms. The possibility of orientation is theoretically predicted either in coherent control schemes using two-color phase-locked laser excitation [21] or in the study of the molecular dynamical response to external fields through symmetry breaking effects by combining the laser fundamental frequency ω and its second harmonic 2ω, where 2ω is resonant with a vibrational transition [22]. More recently we have shown that experimentally available half-cycle pulses (HCP) [23], presenting a unipolar large-amplitude, short-duration component followed by an opposite polarity negative weak-amplitude tail, can also orient polar molecules by creating a coherent superposition of rotational eigenstates [24]. The orientation obtained with this technique turns out to be one of the most efficient. The mechanism rests upon a kick imparted to the molecule by the large-amplitude unipolar component of the pulse imposing its unidirectionality to the molecular system. The weak-amplitude component is shown to have a rather limited effect on the dynamics, basically due to its adiabatic behavior with respect to the rotational motion.

Some optimal control schemes have also been proposed for molecular orientation. In particular, it has been shown that a properly tailored microwave pulse offers, when a reasonable peak power is reached, the possibility to orient, in a temporally recurrent way, a polar diatomic molecule initially in its isotropic $J = M = 0$ state [25, 26]. More recently, tailored laser pulses have been suggested as possible tools in orientation control scenarios [27]. In this article, we are aiming at such a pulse tailoring to achieve orientation by referring to a genetic algorithm [28] dealing with a number of individual laser pulses both characterized by a set of parameters describing their frequency, intensity, phase, and temporal shape. In a previous study [29], we found that the best pulse (for the most efficient orientation), resulting from this type of (physically speaking) black box calculation, is one which imparts a short-duration kick transferring angular momentum to the molecule. This is precisely what we had previously obtained using HCPs [24]. The most efficient orientation, taking into account the molecular response time to the sudden excitation, being reached only a certain time after the pulse if turned off, leads to an important additional bonus : the revival structures of the rotational wave packets that result into predictable, periodic reorientation effects. Actually, in field-free situations, the wave packet dynamics follows the rotational periodicity leading to phase rearrangements such as to produce the same orientation within each rotational period [30].

Optimal control proceeds through an evaluation function, which is taken to be the expectation value of the cosine of the angle between the molecular axis and the laser polarization vector, towards a target. We are interested in long-duration post-pulse field-free orientation which can be reached with short, sudden pulses (in contrast to other studies [13, 31, 32] addressing alignment or orientation during an adiabatic pulse). Even within such an assumably well-delimited frame, the criteria defining the target, which

actually is a rather complex dynamical observable (in opposition to the obtainment of a defined quantum state, for instance), are numerous. It turns out that the outcome of the optimization is very sensitive to any particular choice of a given criterion.

The aim of the present paper is to proceed to a thorough analysis of different criteria and their respective merits in leading to the most efficient orientation with the longest duration. The molecular system that is taken as an illustration is LiF within a rigid rotor approximation. The radiative coupling is described at the dipole approximation level by retaining only the permanent dipole moment and the polarizability in the interaction term. Section 5.2 addresses topics concerning the implementation of the genetic algorithm for the specific problem at hand ; namely, the target (i.e., the orientation criteria to be optimized) and the evaluation function (i.e., a measure of orientation involving its evaluation through numerical solutions of the time-dependent Schrödinger equation [TDSE]). The results are presented in Sec. 5.3, both for calculations at $T = 0$ K (isotropic $J = M = 0$ initial state) and $T = 5$ K.

5.2 Theory

5.2.1 Model

We investigate the laser-induced orientation dynamics of LiF in its ground electronic state, neglecting all internal vibrational motion (the stretching mode is frozen). In addition to this rigid-rotor approximation, we consider the radiative interaction within the dipole approximation, retaining both the permanent dipole moment μ_0 and the polarizability components $\alpha_\parallel$ and $\alpha_\perp$ of LiF at the equilibrium internuclear distance. The complete molecule-plus-field Hamiltonian for this model is

$$\begin{aligned} \hat{H} &= B\hat{J}^2 - \mu_0 \mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left(\alpha_\parallel \cos^2\theta + \alpha_\perp \sin^2\theta\right) \\ &= B\hat{J}^2 - \mu_0 \mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left(\Delta\alpha \cos^2\theta + \alpha_\perp\right), \end{aligned} \tag{5.1}$$

where $\hat{J}^2$ is the angular momentum operator, B the rotational constant, θ the polar angle between the electric field vector $\boldsymbol{\mathcal{E}}(t)$ of the linearly polarized laser field and the molecular axis, and $\Delta\alpha = \alpha_\parallel - \alpha_\perp$. The parameters, obtained from a quantum chemistry calculation [33] (at the MP2 level of perturbation with a 6-311++G(3df,3pd) basis set), are taken as $\mu_0 = 2.5933$ a.u., $\alpha_\parallel = 9.061$ a.u., $\alpha_\perp = 9.218$ a.u., and $B = 5.9173\times10^{-6}$ a.u., giving a rotational period $T_{\rm rot} = h/2B \approx 12.84$ ps. The numerical solution of the TDSE for this model is obtained by developing the wave function $\psi(\theta, \varphi, t)$ of the system on a finite basis set of spherical harmonics,

$$\psi(\theta, \varphi, t) = \sum_{J=0}^{J_{\max}} \sum_{M=-J}^{J} c_{J,M}(t) Y_{J,M}(\theta, \varphi), \tag{5.2}$$

where the $Y_{J,M}(\theta,\varphi)$ are eigenfunctions of $\hat{J}^2$, φ being the azimuthal angle. The TDSE can then be recast as an ensemble of coupled equations of the coefficients $c_{J,M}(t)$, i.e.,

$$\begin{aligned} i\hbar\frac{dc_{J,M}(t)}{dt} &= \left[BJ\left(J+1\right)-\frac{\alpha_{\perp}\mathcal{E}^2(t)}{2}\right]c_{J,M}(t)-\mu_0\mathcal{E}(t)\sum_{J'=0}^{J_{\max}}c_{J',M}(t)\left\langle J,M\right|\cos\theta\left|J',M\right\rangle \\ &\quad -\frac{\Delta\alpha\mathcal{E}^2(t)}{2}\sum_{J'=0}^{J_{\max}}c_{J',M}(t)\left\langle J,M\right|\cos^2\theta\left|J',M\right\rangle, \end{aligned} \tag{5.3}$$

where we have used the notation $\langle\theta,\varphi\,|J,M\rangle = Y_{J,M}(\theta,\varphi)$ and taken into account that M, the projection of the total angular momentum on the field polarization axis, is a good quantum number. Analytical expressions for the scalar products appearing in Eqs. (5.3) are obtained by noticing that $\cos\theta$ and $\cos^2\theta$ can themselves be written in terms of spherical harmonics, resulting in the integration of a product of three spherical harmonics [34], yielding

$$\left\langle J,M\right|\cos\theta\left|J+1,M\right\rangle = \left[\frac{\left(J+M+1\right)\left(J-M+1\right)}{\left(2J+3\right)\left(2J+1\right)}\right]^{1/2} \tag{5.4a}$$

$$\left\langle J,M\right|\cos\theta\left|J-1,M\right\rangle = \left[\frac{\left(J+M\right)\left(J-M\right)}{\left(2J+1\right)\left(2J-1\right)}\right]^{1/2} \tag{5.4b}$$

$$\left\langle J,M\right|\cos\theta\left|J',M\right\rangle = 0 \text{ for } |J'-J|\neq\pm1 \tag{5.4c}$$

and (see also Ref. [35])

$$\left\langle J,M\right|\cos^2\theta\left|J+2,M\right\rangle = \frac{1}{2J+3}\left[\frac{\left(J+M+2\right)\left(J+M+1\right)\left(J-M+2\right)\left(J-M+1\right)}{\left(2J+5\right)\left(2J+1\right)}\right]^{1/2} \tag{5.5a}$$

$$\left\langle J,M\right|\cos^2\theta\left|J-2,M\right\rangle = \frac{1}{2J-1}\left[\frac{\left(J+M\right)\left(J+M-1\right)\left(J-M\right)\left(J-M-1\right)}{\left(2J+1\right)\left(2J-3\right)}\right]^{1/2} \tag{5.5b}$$

$$\left\langle J,M\right|\cos^2\theta\left|J,M\right\rangle = \frac{1}{3}+\frac{2}{3}\left[\frac{J\left(J+1\right)-3M^2}{\left(2J+3\right)\left(2J-1\right)}\right] \tag{5.5c}$$

$$\left\langle J,M\right|\cos^2\theta\left|J',M\right\rangle = 0 \text{ for } |J'-J|\neq0,\pm2. \tag{5.5d}$$

Equations (5.3) are solved using a 4th-order Runge-Kutta propagator [36], with the initial state (at $t=0$) taken to be the ground rotational state $J=M=0$. Such a state, which could be prepared by laser cooling methods [37], corresponds to an isotropic distribution and offers the advantage of rotational excitations only to high-J, $M=0$ manifolds with the molecular axis nearly parallel to the field (i.e., alignment). The measure of the orientation, which will be taken as an evaluation function for the genetic algorithm, is the expectation value of $\cos\theta$ [4],

$$\left\langle\cos\theta\right\rangle(t) = \int_0^{2\pi}\int_0^{\pi}\left|\psi(\theta,\varphi;t)\right|^2\cos\theta\sin\theta d\theta d\varphi, \tag{5.6}$$

calculated from the $c_{J,M}(t)$ coefficients by use of Eqs. (5.4). Roughly speaking, orientation is achieved for large absolute values of $\langle\cos\theta\rangle$. More precise criteria, defining the target of the genetic algorithm, will be described later.

5.2.2 Description of the laser field

The electromagnetic field, to be adjusted by the optimization procedure, is modeled by a sum of individual linearly-polarized pulses,

$$\mathcal{E}(t) = \sum_{n=1}^{N} \mathcal{E}_n(t) \sin\left(\omega_n t + \phi_n\right). \tag{5.7}$$

The electric field amplitude envelope functions $\mathcal{E}_n(t)$ are given sine-squared forms for the switching on and off regimes, with a constant plateau value in between,

$$\mathcal{E}_n(t) = \begin{cases} 0 & \text{if} \quad t \leq t_{n0} \\ \mathcal{E}_{n0} \sin^2\left[\frac{\pi}{2}\left(\frac{t-t_{n0}}{t_{n1}-t_{n0}}\right)\right] & \text{if} \quad t_{n0} \leq t \leq t_{n1} \\ \mathcal{E}_{n0} & \text{if} \quad t_{n1} \leq t \leq t_{n2} \\ \mathcal{E}_{n0} \sin^2\left[\frac{\pi}{2}\left(\frac{t_{n3}-t}{t_{n3}-t_{n2}}\right)\right] & \text{if} \quad t_{n2} \leq t \leq t_{n3} \\ 0 & \text{if} \quad t \geq t_{n3} \end{cases} \tag{5.8}$$

Each pulse is thus characterized by a set of 7 free parameters, namely its frequency ω_n, absolute phase ϕ_n, and maximum field amplitude $\mathcal{E}_{n0}$, together with 4 positive times characterizing its shape (i.e., its time origin t_{n0}, rise time $t_{n1} - t_{n0}$, plateau duration $t_{n2} - t_{n1}$, and extinction time $t_{n3} - t_{n2}$). It is worth noting that such a parameterization is one of many ways of reducing a complicated laser pulse to a limited number of parameters for the optimization algorithm. This particular choice can facilitate the extraction of information from the pulse, to better understand the physical processes involved during the laser-molecule interaction [29, 38]. In experiments, the superposition of time-delayed individual pulses would presumably not be the most convenient procedure, but the resulting total field $\mathcal{E}(t)$ could be obtained by modern pulse shaping techniques [39–41].

5.2.3 Optimization methodology

How to tailor an electromagnetic field, within the model described by Eq. (5.7), to reach the best possible orientation is a parameter optimization problem involving, as we have previously stated, $7N - 1$ *variables* (N being the total number of individual pulses, the minus one appearing because the origin of time is arbitrarily set). A dynamical measure of orientation is provided by $\langle \cos\theta \rangle (t)$, Eq. (5.6). The *evaluation function* is thus the algorithm calculating the ingredients of Eq. (5.6), basically the wave packet at time t, through the numerical evaluation of the time-dependent coupled equations (5.3). The optimization procedure implies the maximization (or minimization) of a single value, the *target* or *criterion*, that has to be extracted from the complex orientation dynamics resulting from the application of a given laser field $\mathcal{E}(t)$. Different criteria present different advantages (and/or restrictions) and may even lead to different electromagnetic fields resulting into different orientation dynamics. We are putting them, for convenience, into two groups.

(i) The first group gathers simple criteria, emphasizing either the efficiency or the duration of the orientation. From previous studies [22], we know the difficulty of orienting

molecules during the time where the laser pulse is on, and expect that the orientation (even of low efficiency) obtained at the end of the pulse to be determinant for the subsequent dynamics. As a consequence, a criterion can be selected that searches for the maximum of the absolute expectation value of $\cos\theta$ at a time t_f taken as the maximum time allowed for all individual pulses before switching off ($t_{3n} \leq t_f$). This is precisely the choice that is investigated in Ref. [29] and leads to the kick mechanism. However, more flexibility can be gained by searching for the same maximum, but for any time within the interval $[t_f, t_f + T_{\rm rot}]$, i.e., for an entire rotational period after the interaction with the laser pulse,

$$j_1 \doteqdot \max_{t\in[t_f,t_f+T_{\rm rot}]} \left|\langle\cos\theta\rangle\,(t)\right|. \tag{5.9}$$

The addition of $T_{\rm rot}$ is to fully take advantage of the recursive behavior (i.e., revival structure) of the field-free orientation dynamics after the molecular rotational response to the electromagnetic field. Other criteria may put the emphasis on the duration of the orientation, rather than its efficiency. A possible choice is then

$$j_2 \doteqdot \max_{t\in[t_f,t_f+T_{\rm rot}]} \frac{\tau}{T_{\rm rot}}, \tag{5.10}$$

where τ designates a duration of orientation (i.e., a time interval over which $\langle\cos\theta\rangle$ remains relatively high). To be more specific, τ can be taken as the total duration of all time intervals, for which

$$\frac{j_1}{\sqrt{2}} \leq \left|\langle\cos\theta\rangle\,(t)\right| \leq j_1 \tag{5.11}$$

with $t \in [t_f, t_f + T_{\rm rot}]$. The weakness of this criterion is that, if taken alone, it may not lead to any orientation when the value of j_1 is small.

(ii) Physically more sound criteria, that combine the advantages of the previous requirements and search for a compromise between the efficiency and the duration of the orientation, are gathered in the second group. Such a compromise is well accounted for by a carefully built hybrid criterion given as

$$j_3 \doteqdot \max_{t\in[t_f,t_f+T_{\rm rot}]} \left[\left|\langle\cos\theta\rangle\,(t)\right| + \frac{\tau}{T_{\rm rot}} - \left|\left|\langle\cos\theta\rangle\,(t)\right| - \frac{\tau}{T_{\rm rot}}\right|\right], \tag{5.12}$$

that implies simultaneously the maximization of $|\langle\cos\theta\rangle\,(t)|$ and of $\tau/T_{\rm rot}$, precisely as j_1 and j_2 would have done. The additional absolute value term in Eq. (5.12) is to properly balance the relative contributions of the two criteria j_1 and j_2 (both with values in the $[0, 1]$ interval), resulting into a lower value for j_3 when both criteria are simultaneously satisfied.

A different and dynamically more significant criterion, and presumably good candidate for an efficiency versus duration compromise, would be the maximization of the time average of $\langle\cos\theta\rangle$ over a rotational period, i.e.,

$$j \doteqdot \max \left|\frac{1}{T_{\rm rot}} \int_{t_f}^{t_f+T_{\rm rot}} \langle\cos\theta\rangle\,(t)dt\right|. \tag{5.13}$$

But, due to symmetry and periodicity properties of $\langle \cos\theta \rangle (t)$, such an integral is strictly zero (a detailed proof is given in Appendix). This means that it is impossible to create a superposition of rotational states leading to the orientation, on average over an entire rotational period, of a quantum rotor. Keeping basically this idea in mind, similar criteria taking advantage of time-averaged dynamics can be considered, such as

$$j_4 \doteqdot \max \frac{1}{T_{\rm rot}} \int_{t_f}^{t_f+T_{\rm rot}} \langle \cos\theta \rangle^2 (t) dt \tag{5.14}$$

or, even more flexible and more sophisticated,

$$j_5 \doteqdot \max \frac{1}{T_{\rm rot}} \int_{t_f}^{t_f+T_{\rm rot}} \mathcal{C}^2(t) dt, \tag{5.15}$$

where

$$\mathcal{C}(t) = \begin{cases} 0.1 \langle \cos\theta \rangle (t) & \text{if } \langle \cos\theta \rangle (t) < 0.4 \\ \langle \cos\theta \rangle (t) & \text{elsewhere} \end{cases} \tag{5.16}$$

is tailored to put the emphasis on time intervals where $\langle \cos\theta \rangle (t)$ is greater than some fixed value (0.4 in this example), by reducing the weight at other times by some convenient factor (0.1 in this example). These specific values could in turn be subject to tweaking.

The target (optimization criterion) being specified, basically three families of methods are available for the optimization process itself : either purely deterministic, purely stochastic, or hybrid. Each offers complementary advantages within its own limitations. Among deterministic methods, the non-linear conjugated gradient is one of the more rapidly convergent, but presents the risk of remaining trapped in a local extremum [42]. Stochastic methods do not present such a drawback and are not, in principle, sensitive to the number of local extrema. But, at zeroth-order (referring to the only information concerning the values of the evaluation function at each point of the parameter space) these methods may require much memory and be presumably too slow due to the large number of iterations required. We have taken full advantage of the possibilities of the second family by implementing an evolutionary algorithm based on a classical genetic algorithm with a floating point representation [43].

Tentatives have also been made to construct a hybrid algorithm by creating a mutation-by-gradient operator. In this approach, every individual is mutated by the application of a few iterations of a conjugated gradient algorithm [44]. Unfortunately, this does not lead to any noticeable decrease of the convergence time [38]. In the following, only results obtained with the purely stochastic approach are presented.

5.3 Results

For the sake of simplicity, all calculations (unless otherwise specified) are conducted using $N = 2$ lasers. Moreover, their basic parameters are confined within the following

limits :

$$\begin{aligned} \mathcal{I}_n &\in \left[10^8, 3\times10^{13}\right] \text{ W/cm}^2 \\ \omega_n &\in [500, 4000] \text{ cm}^{-1} \\ \phi_n &\in [0, 2\pi] \\ t_{n3} &\leq \frac{T_{\text{rot}}}{10} \end{aligned}$$

The electric field $\mathcal{E}_{n0}$ used in Eq. (5.8) is related to the intensity $\mathcal{I}_n$ through the relation $\mathcal{E}_{n0} = \sqrt{2\mathcal{I}_n/(\varepsilon_0 c)}$, with ε_0 the electric constant and c the speed of light. The ionization potential of LiF is 12.9 eV [45], and the first ionization intensity threshold is predicted, from tunneling ionization models [46], to be 1.1×10^{14} W/cm^2 in the IR region. The current limit thus ensures that no ionization can occur. In the genetic algorithm, each generation is made up of ten individuals and convergence is typically achieved after 200 iterations.

The optimization results are presented in two paragraphs. Section 5.3.1 is devoted to the comparative analysis of the most relevant criteria leading to an optimized field together with the orientation dynamics, without any consideration of temperature effects (the initial state being taken as $J = M = 0$ as it would be for 0 K). Section 5.3.2 deals with the very first optimal control of orientation taking properly into account initial thermal distributions (at 5 K), using some of the criteria defined in Sec. 5.3.1.

5.3.1 Comparative analysis of the criteria ($T = 0$ K)

5.3.1.1 Simple criteria

Figure 5.1 displays the results obtained by taking as a target j_1 [Eq. (5.9)], aiming at the most efficient orientation (without consideration of its duration). The inset gives the optimal field shape corresponding to a sudden (less than 1.3 ps duration), intense (3×10^{13} W/cm^2) pulse. The individual pulses, with comparable intensities and frequencies in a ratio of 2, are responsible for the double wiggles of the overall electric field amplitude. Although no orientation effect is observed during the laser pulse, as expected for a sudden pulse [24], the time-dependent behavior of $\langle\cos\theta\rangle$ shows excellent orientation efficiency with a value -0.79 reached at about $t = 7.15$ ps. As this orientation is achieved after the pulse is off, it recurrently occurs with the rotational periodicity of the molecule. These recurrences in the post-pulse dynamics of $\langle\cos\theta\rangle$ are common to all the present calculations. Unfortunately, this simple criterion fails to produce orientation lasting a long time. τ, defined as the time duration over which an orientation exceeding $j_1/\sqrt{2}$ is kept, is less than 0.39 ps.

Figure 5.2 gives the dynamics resulting from target j_2 [Eq. (5.10)], aiming at the longest duration of orientation (without consideration of its efficiency). The electric field (given in the inset) is built up from two intense, sudden, high frequency and well separated (sequential) pulses. As expected, the value of orientation duration is large, $\tau \approx 3.38$ ps, but more disappointingly the orientation is very inefficient, with the maximum value of $|\langle\cos\theta\rangle|$ reaching only 0.001. These two examples show the limitations of simple criteria

in producing physically sound orientation effects and urge for more sophisticated hybrid criteria.

5.3.1.2 Hybrid criteria

The results, when using target j_3 [Eq. (5.12)] are displayed in Fig. 5.3. The optimized laser field (given in the inset) is very sudden, intense, and high frequency. Here again, although the duration is quite long, $\tau \approx 3$ ps, the orientation achieved remains small, with $|\langle\cos\theta\rangle| < 0.13$. It is worthwhile noting that the maximum duration which is actually retained by the calculation is less than the one adopted as the upper limit ($t_{n3} \leq T_{\rm rot}/10 \approx 1.28$ ps). Thus, the optimization technique is not biased by such a choice that arbitrarily would impose a sudden pulse. This turns out to be a common feature in the present study (except for the case of criterion j_1, see Fig. 5.1).

The best results are obtained for the hybrid criteria j_4 [Eq. (5.14)] and j_5 [Eq. (5.15)], combining efficiency and duration through a time-average of $\langle\cos\theta\rangle^2$ over a rotational period. They are gathered in Fig. 5.4. Panel (a) illustrates the dynamics under criterion j_4 induced by an intense laser field, lasting approximately 1 ps with the double wiggling structure due to the overlapping of individual components with frequencies in a ratio of 2. The maximum orientation $\langle\cos\theta\rangle = -0.624$ occurs at about $t = 6.95$ ps and lasts for $\tau = 0.36$ ps. These performances are comparable to what is obtained by using the simple criterion j_1. But an improvement seems possible with the more flexible target j_5, as shown in Fig. 5.4(b). The laser field is very sudden (less than 0.5 ps duration), intense (3×10^{13} W/cm^2), and built up of two approximately same-frequency components, giving the beat-like structure. Although this shape is similar to the one that leads to a kick mechanism [29], the rotational dynamics in this case are more complicated and cannot be reduced to the action of one or two sudden impacts on the molecule. The compromise between efficiency and duration which is achieved is very satisfactory and leads to one of the very best results in laser control of orientation ($\langle\cos\theta\rangle = -0.679$; $\tau \approx 0.69$ ps).

This result could, presumably, be improved further by refining the parameters involved in the definition of $\mathcal{C}(t)$, Eq. (5.16), but still remains of rather short duration for a full stereodynamical control of bimolecular reactive collisions. Two arguments could however be provided to support the quality of the present achievement. The first is to consider orientation as a tool for controlling more rapid processes, such as half-collisions. Isotope separation in the photodissociation of HD^+ is such an example [6]. The second is related to the revival structure : re-orientation at predictable times within each rotational period may be made use of in some adequately synchronized full collision process.

5.3.2 Orientation control under thermal averaging

Previous studies of laser-induced alignment or orientation have shown that the degree of alignment or orientation obtained decreases rapidly as the initial rotational temperature to which the pulse is applied increases [29, 35, 47–49]. The orientation is actually erased when Boltzmann averaging the rotational distributions due to the fact that a pulse optimized for the $J = 0$, $M = 0$ state is no more appropriate for orienting rotationally excited

initial states. The average to be performed is given by

$$\langle\langle \cos\theta \rangle\rangle (t) = Q^{-1} \sum_{J}^{J_{\max}} \exp\left[\frac{-BJ(J+1)}{k_B T}\right] \sum_{M=-J}^{J} \langle \cos\theta \rangle_{J,M} (t), \tag{5.17}$$

where Q is the partition function

$$Q = \sum_{J}^{J_{\max}} (2J+1) \exp\left[\frac{-BJ(J+1)}{k_B T}\right], \tag{5.18}$$

k_B and T being the Boltzmann constant and the rotational temperature, respectively. $\langle \cos\theta \rangle_{J,M} (t)$ describes the orientation dynamics for a given single initial state J, M calculated as previously through Eq. (5.6) by an exact wave packet propagation performed using Eq. (5.3) with $|J, M\rangle$ as the initial rotational state.

The upper panel (a) of Fig. 5.5 displays the thermally averaged orientation dynamics as resulting from the simple criterion j_1 using again two laser pulses in Eq. (5.7). The optimization is carried out on the evaluation function given by Eq. (5.17) (solid line) instead of the one given by Eq. (5.6) (dotted line) for comparison. The rather fast decrease of orientation, even for temperatures not exceeding 5 K, can be observed from the dotted curve showing the dynamics of $\langle\langle \cos\theta \rangle\rangle (t)$ with the field optimized for $J = M = 0$ (see inset of Fig. 5.1). Namely, the efficiency is divided by a factor larger than 5 (when compared to $\langle \cos\theta \rangle$ in Fig. 5.1). Such an effect is basically expected since a thermally averaged rotational population not only contains a wider distribution of J's but also of M's. The latter, which can in any way not be decreased using a linearly polarized laser, is responsible for the fast damping of orientation effects. The optimally tailored field, because it takes precisely into account the initial thermal averaged rotational population in the evaluation function of the optimal control scheme, can restitute acceptable orientation, with an efficiency $\langle\langle \cos\theta \rangle\rangle = -0.27$ and a duration $\tau = 0.5$ ps. Further improvement restituting the structures in $\langle\langle \cos\theta \rangle\rangle (t)$ can be obtained by increasing the flexibility of the optimization algorithm's search space using additional individual pulses in Eq. (5.7). The result with $N = 3$ is displayed in Fig. 5.5(b). An orientation efficiency of $\langle\langle \cos\theta \rangle\rangle = -0.38$ is achieved with a duration of $\tau = 0.26$ ps. We have also checked the performances that could be reached using a hybrid criterion. Figure 5.6 displays the dynamics of $\langle\langle \cos\theta \rangle\rangle (t)$ after optimization of a field, built up from $N = 3$ individual pulses, according to criterion j_4. This leads to the best efficiency/duration compromise, with $\langle\langle \cos\theta \rangle\rangle = -0.30$ and $\tau \approx 0.34$ ps.

5.4 Conclusion

One of the basic requirements of optimal control schemes is the definition of the target, not only involving an evaluation function that has to be calculated within reasonable time limits, but more importantly reflecting a clear and sound characteristic of the physical process that is under control. Some processes, involving the achievement of a given

single quantum state, for instance, are simpler in this respect. Others, like alignment or orientation, are much more complicated because they refer to an observable involving a superposition of a manifold of quantum states within a whole dynamical evolution. It is therefore expected that several different targets could be conceived, with limited possibility of inferring on their respective merits prior to the calculation.

Among the large variety of possible criteria that we have checked, the most meaningful ones are collected in this work and classified into two groups : *simple*, when they deal either with the maximum efficiency of orientation or with its maximum duration ; *hybrid*, when they combine the two requirements into a compromise. Excellent orientation is obtained from control scenarios starting, as an initial state, from an isotropic geometry (i.e., $J = M = 0$, the case of $T = 0$ K), with $|\langle \cos\theta \rangle|$ reaching a maximum value close to 0.7 over a time interval of about 0.7 ps. Moreover, such a result is achieved using only two individual laser pulses, having in common their high intensity (of the order of 10^{13} W/cm^2) and short duration (less than 1.28 ps) as characteristic features.

Increasing the temperature has as consequence the rotational excitation of the molecule with increasing J's and M's levels being populated. This is at the origin of the orientation damping that is currently observed. A possible way to overcome this damping is to optimally tailor a laser pulse taking into account the thermal averaging in the target itself. This is actually what has been done and leads to rather good orientation achieved by superposing three individual laser pulses (efficiency of about 0.30 for a duration of 0.34 ps at $T = 5$ K). This encouraging result opens as a prospective research, the way to the ellipticity control of non linearly-polarized lasers [50], not only aiming at the orientation of rotationally hot systems, but also at the 3D alignment/orientation of polyatomic molecules [51]. Work along these lines is actively pursued in our group.

Acknowledgments: Financial support from the French Ministry of Research through an *Action Concertée Incitative Jeunes Chercheurs* is gratefully acknowledged.

Appendix : Time-average of $\langle \cos\theta \rangle$

The wave function of a free rigid rotor (in a single quantum state M) can be expanded at any time on a basis set of spherical harmonics as

$$\begin{aligned} \psi(\theta,\varphi,t) &= \sum_{J=0}^{\infty} c_J(t) Y_{J,M}(\theta,\varphi) \\ &= \sum_{J=0}^{\infty} \tilde{c}_J e^{-iBJ(J+1)t/\hbar} Y_{J,M}(\theta,\varphi), \end{aligned} \tag{5.19}$$

where the $\tilde{c}_J$'s are time-invariant complex coefficients. Using this result in Eq. (5.6), we have

$$\begin{aligned}\langle\cos\theta\rangle(t) &= \int_0^{2\pi}\int_0^{\pi}\left[\sum_{J=0}^{\infty}\tilde{c}_J^* e^{iBJ(J+1)t/\hbar}Y_{J,M}^*(\theta,\varphi)\right]\left[\sum_{J'=0}^{\infty}\tilde{c}_{J'}e^{-iBJ'(J'+1)t/\hbar}Y_{J',M}(\theta,\varphi)\right]\\ &\quad\times\cos\theta\sin\theta d\theta d\varphi\\ &= \sum_{J=0}^{\infty}\sum_{J'=0}^{\infty}\tilde{c}_J^*\tilde{c}_{J'}e^{iB[J(J+1)-J'(J'+1)]t/\hbar}\\ &\quad\times\int_0^{2\pi}\int_0^{\pi}Y_{J,M}^*(\theta,\varphi)Y_{J',M}(\theta,\varphi)\cos\theta\sin\theta d\theta d\varphi. \end{aligned}\tag{5.20}$$

Considering Eqs. (5.4), the sums are reduced to

$$\begin{aligned}\langle\cos\theta\rangle(t) &= \sum_{J=0}^{\infty}\tilde{c}_J^*\tilde{c}_{J+1}e^{-i2B(J+1)t/\hbar}\langle J,M|\cos\theta|J+1,M\rangle\\ &\quad+\sum_{J=1}^{\infty}\tilde{c}_J^*\tilde{c}_{J-1}e^{i2BJt/\hbar}\langle J,M|\cos\theta|J-1,M\rangle. \end{aligned}\tag{5.21}$$

The only time-dependent term in Eq. (5.21) being the exponential, the calculation of the time-average of $\langle\cos\theta\rangle$ rests on an integral of the form

$$\begin{aligned}\int_t^{t+T_{\rm rot}}e^{i2BJt'/\hbar}dt' &= \frac{i\hbar}{2BJ}e^{i2BJt'/\hbar}\Big|_{t'=t}^{t'=t+T_{\rm rot}}\\ &= \frac{i\hbar}{2BJ}e^{i2BJt/\hbar}\left[e^{i2BJT_{\rm rot}/\hbar}-1\right], \end{aligned}\tag{5.22}$$

which, considering that $T_{\rm rot}=h/2B=\pi\hbar/B$, results in

$$\int_t^{t+T_{\rm rot}}e^{i2BJt'/\hbar}dt'=\frac{i\hbar}{2BJ}e^{i2BJt/\hbar}\left[e^{i2J\pi}-1\right]=0.\tag{5.23}$$

The criterion given in Eq. (5.13), is thus strictly nil. Moreover, this means that it is impossible to create a superposition of rotational states such that, on average over time, a molecule is more oriented in a given direction [the proof is easily extended to the case where the sum in Eq. (5.19) runs over both J's and M's].

References

[1] K. H. Kramer and R. B. Bernstein, J. Chem. Phys. **42**, 767 (1965).

[2] T. D. Hain, R. M. Moision, and T. J. Curtiss, J. Chem. Phys. **111**, 6797 (1999).

[3] H. J. Loesch and A. Remscheid, J. Chem. Phys. **93**, 4779 (1990).

[4] B. Friedrich and D. R. Herschbach, Z. Phys. D **18**, 153 (1991).

[5] M. G. Tenner, E. W. Kuipers, A. W. Kleyn, and S. Stolte, J. Chem. Phys. **94**, 5197 (1991).

[6] E. Charron, A. Giusti-Suzor, and F. H. Mies, Phys. Rev. A **49**, R641 (1994).

[7] H. Stapelfeldt, H. Sakai, E. Constant, and P. B. Corkum, Phys. Rev. Lett. **79**, 2787 (1997).

[8] T. Seideman, Phys. Rev. A **56**, R17 (1997).

[9] B. K. Dey, M. Shapiro, and P. Brumer, Phys. Rev. Lett. **85**, 3125 (2000).

[10] M. D. Poulsen, E. Skovsen, and H. Stapelfeldt, J. Chem. Phys. **117**, 2097 (2002).

[11] B. A. Zon and B. G. Katsnel'son, Sov. Phys. JETP **42**, 595 (1975), [Zh. Eksp. Teor. Fiz. **69**, 1166 (1975)].

[12] B. Friedrich and D. Herschbach, Phys. Rev. Lett. **74**, 4623 (1995).

[13] A. Keller, C. M. Dion, and O. Atabek, Phys. Rev. A **61**, 023409 (2000).

[14] C. M. Dion, A. Keller, O. Atabek, and A. D. Bandrauk, Phys. Rev. A **59**, 1382 (1999a).

[15] D. Normand, L. A. Lompré, and C. Cornaggia, J. Phys. B : At. Mol. Opt. Phys. **25**, L497 (1992).

[16] J. H. Posthumus, J. Plumridge, L. J. Frasinski, K. Codling, A. J. Langley, and P. F. Taday, J. Phys. B : At. Mol. Opt. Phys. **31**, L985 (1998).

[17] C. Ellert and P. B. Corkum, Phys. Rev. A **59**, R3170 (1999).

[18] S. Banerjee, G. R. Kumar, and D. Mathur, Phys. Rev. A **60**, R3369 (1999).

[19] H. Sakai, C. P. Safvan, J. J. Larsen, K. M. Hilligs e, K. Hald, and H. Stapelfeldt, J. Chem. Phys. **110**, 10235 (1999).

[20] K. Hoshina, K. Yamanouchi, T. Ohshima, Y. Ose, and H. Todokoro, Chem. Phys. Lett. **353**, 27 (2002).

[21] M. J. J. Vrakking and S. Stolte, Chem. Phys. Lett. **271**, 209 (1997).

[22] C. M. Dion, A. D. Bandrauk, O. Atabek, A. Keller, H. Umeda, and Y. Fujimura, Chem. Phys. Lett. **302**, 215 (1999b).

[23] D. You, R. R. Jones, P. H. Bucksbaum, and D. R. Dykaar, Opt. Lett. **18**, 290 (1993).

[24] C. M. Dion, A. Keller, and O. Atabek, Eur. Phys. J. D **14**, 249 (2001).

[25] R. S. Judson, K. K. Lehmann, H. Rabitz, and W. S. Warren, J. Mol. Spectrosc. **223**, 425 (1990).

[26] J. Ortigoso, Phys. Rev. A **57**, 4592 (1998).

[27] K. Hoki and Y. Fujimura, Chem. Phys. **267**, 187 (2001).

[28] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs* (Springer, Berlin, 1996), 3rd ed.

[29] C. M. Dion, A. Ben Haj-Yedder, E. Cancès, C. Le Bris, A. Keller, and O. Atabek, Phys. Rev. A **65**, 063408 (2002).

[30] T. Seideman, Phys. Rev. Lett. **83**, 4971 (1999).

[31] T. Kanai and H. Sakai, J. Chem. Phys. **115**, 5492 (2001).

[32] S. Guérin, L. P. Yatsenko, H. R. Jauslin, O. Faucher, and B. Lavorel, Phys. Rev. Lett. **88**, 233601 (2002).

[33] *Gaussian 98* (Revision A.9), M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle and J. A. Pople, Gaussian, Inc., Pittsburgh PA, 1998.

[34] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).

[35] J. Ortigoso, M. Rodríguez, M. Gupta, and B. Friedrich, J. Chem. Phys. **110**, 3870 (1999).

[36] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN* (Cambridge University Press, Cambridge, 1992), 2nd ed.

[37] J. T. Bahns, P. L. Gould, and W. C. Stwalley, Adv. At. Mol. Opt. Phys. **42**, 171 (2000).

[38] A. Auger, A. Ben Haj-Yedder, E. Cancès, C. Le Bris, C. M. Dion, A. Keller, and O. Atabek, Math. Models Methods Appl. Sci. **12**, 1281 (2002).

[39] M. R. Fetterman, D. Goswami, D. Keusters, W. Yang, J.-K. Rhee, and W. S. Warren, Opt. Express **3**, 366 (1998).

[40] E. Zeek, K. Maginnis, S. Backus, U. Russek, M. Murnane, G. Mourou, H. Kapteyn, and G. Vdovin, Opt. Lett. **24**, 493 (1999).

[41] T. Brixner, A. Oehrlein, M. Strehle, and G. Gerber, App. Phys. B **70**, S119 (2000).

[42] J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer, Berlin, 1999).

[43] M. Vose, in *Foundations of Genetic Algorithms II*, edited by D. Whitley (Morgan Kaufmann, San Mateo, 1993).

[44] A. Ben Haj-Yedder, in *Automatic Differentiation of Algorithms : From Simulation to Optimization*, edited by G. Corliss, C. Faure, A. Griewank, and L. Hascoet (Springer, Berlin, 2002).

[45] J. C. Pinheiro, M. Trsic, and A. B. F. da Silva, J. Mol. Struct. (THEOCHEM) **539**, 29 (2001).

[46] P. Dietrich, P. B. Corkum, D. T. Strickland, and M. Laberge, in *Molecules in Laser Fields*, edited by A. D. Bandrauk (Dekker, New York, 1994), chap. 4, pp. 181–216.

[47] M. Machholm and N. E. Henriksen, Phys. Rev. Lett. **87**, 193001 (2001).

[48] M. Machholm, J. Chem. Phys. **115**, 10724 (2001).

[49] T. Seideman, J. Chem. Phys. **115**, 5965 (2001).

[50] J.-H. Kim, J.-M. Yuan, W.-K. Liu, and C. H. Nam, Phys. Rev. A **63**, 043420 (2001).

[51] J. J. Larsen, K. Hald, N. Bjerre, H. Stapelfeldt, and T. Seideman, Phys. Rev. Lett. **85**, 2470 (2000).

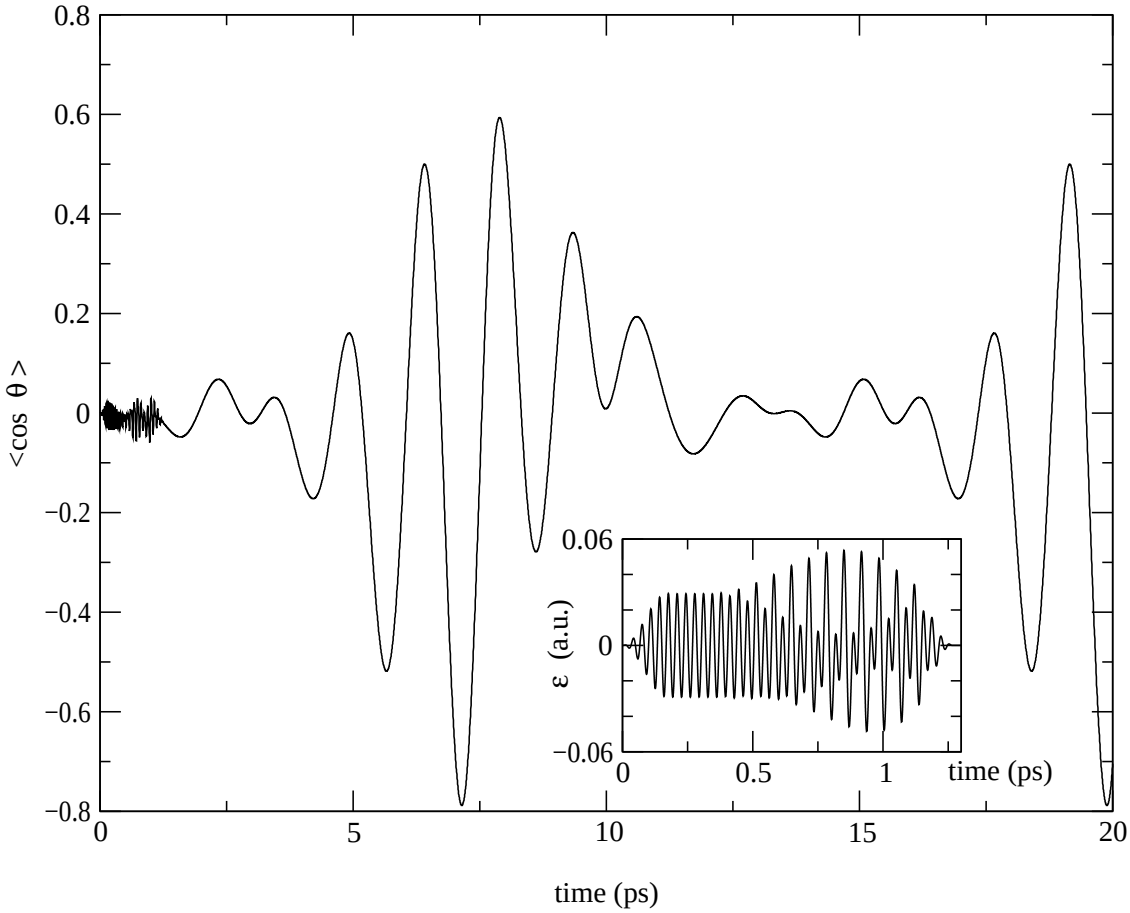


FIG. 5.1 – Orientation dynamics in terms of the time evolution of the expectation value of the cosine of the angle between the LiF molecular axis and the linearly polarized field polarization vector, resulting from an optimization of criterion j_1 (see main text for definition). The inset displays the time evolution of the optimized laser field.

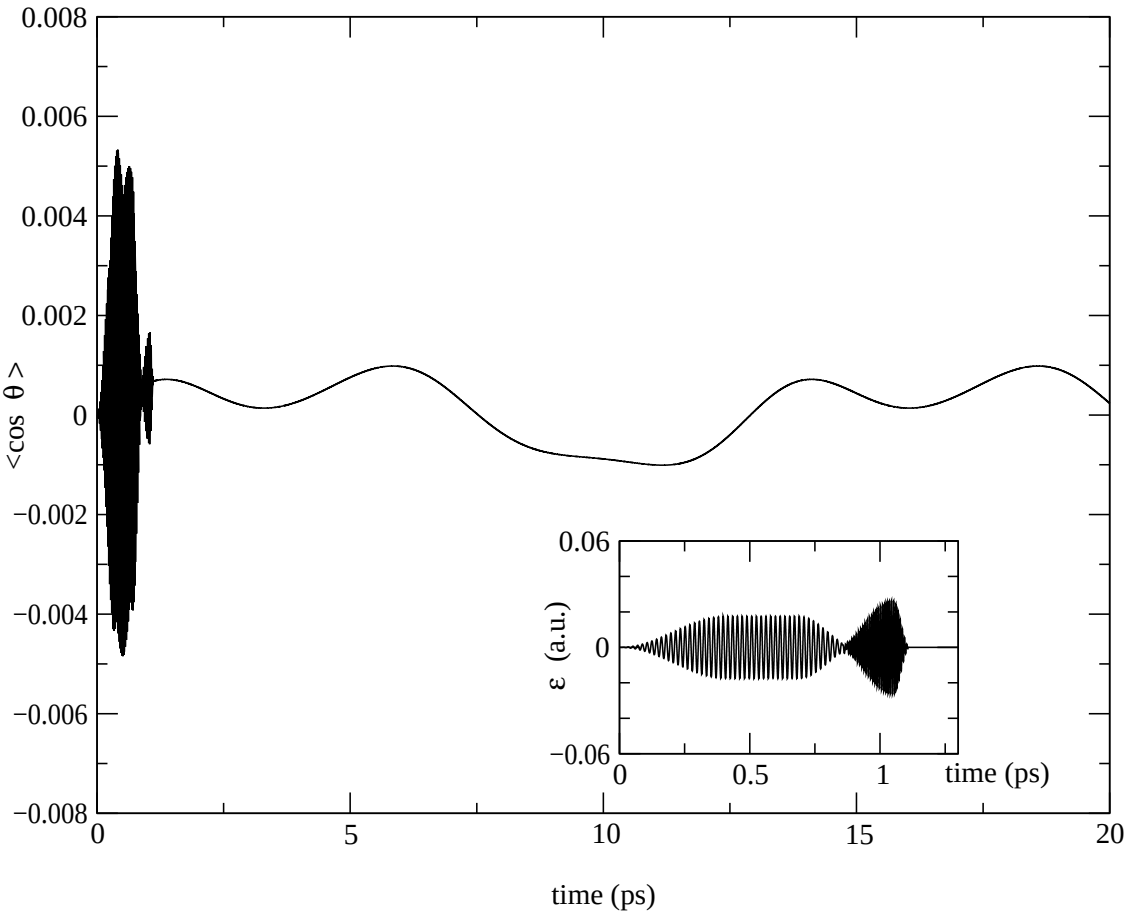


FIG. 5.2 – Same as Fig. 5.1, but with criterion j_2.

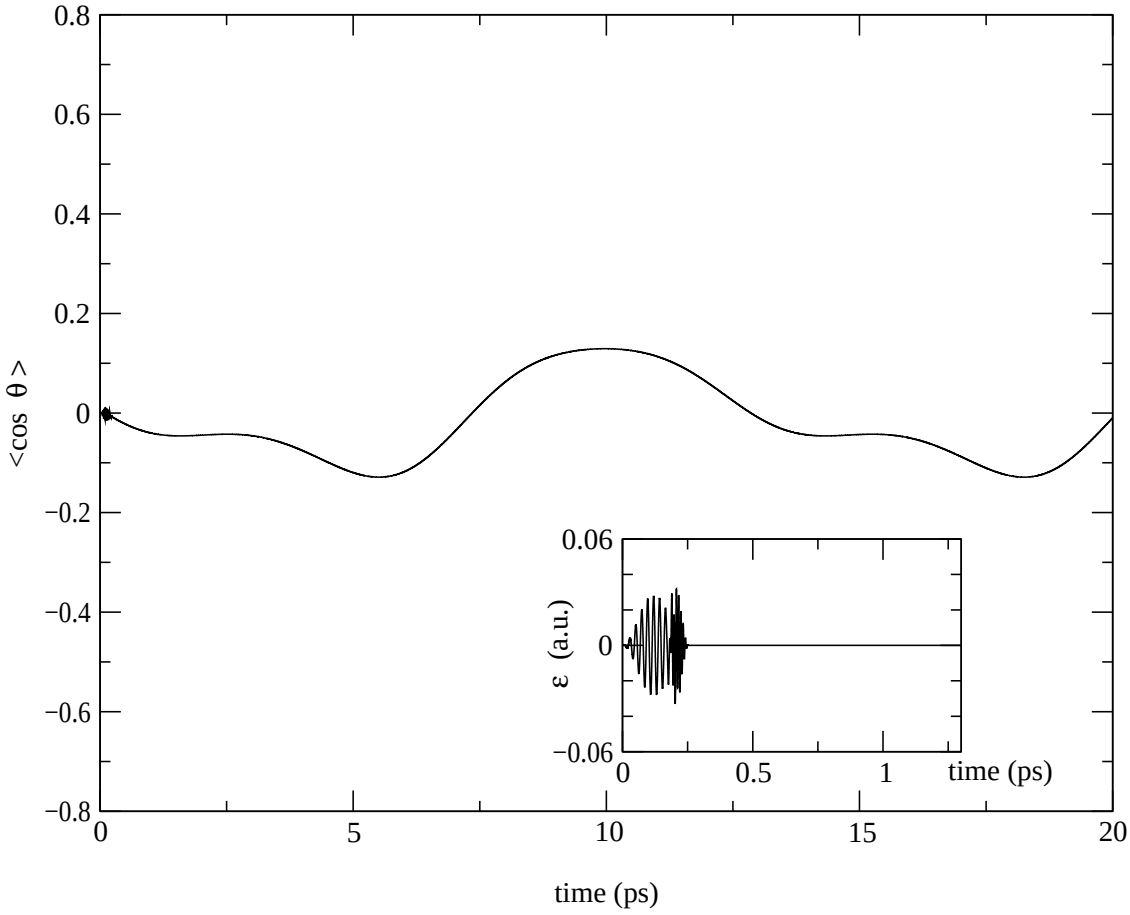


FIG. 5.3 – Same as Fig. 5.1, but with criterion j_3.

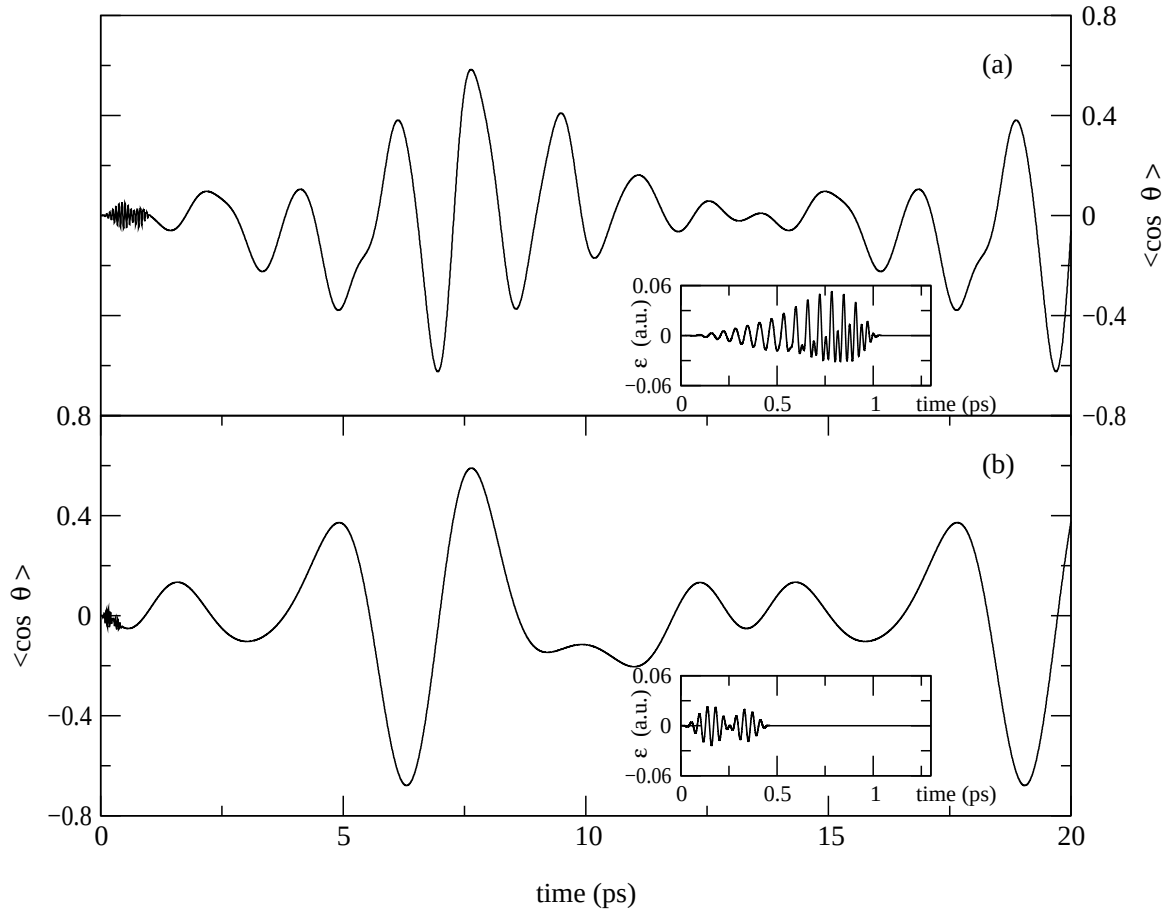


FIG. 5.4 – Same as Fig. 5.1, but with criteria (a) j_4 and (b) j_5.

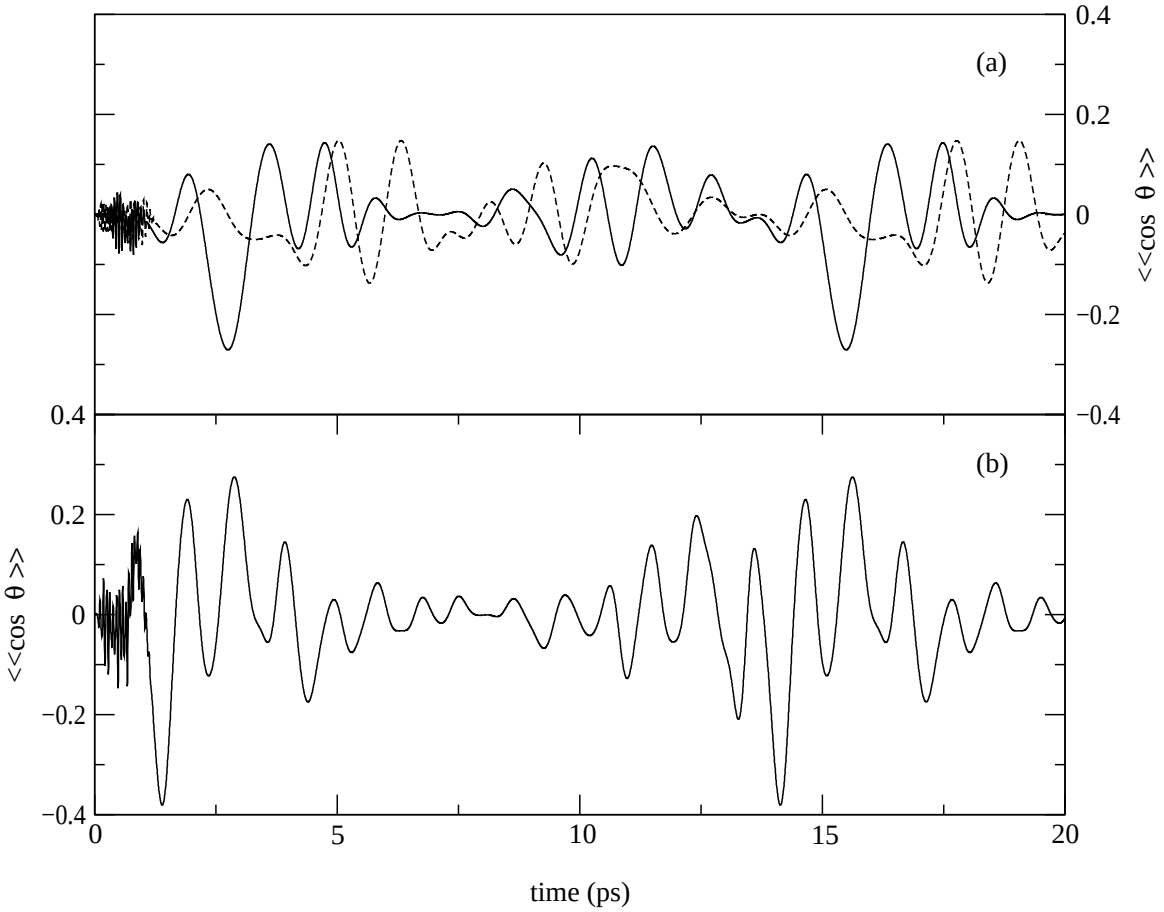


FIG. 5.5 – Orientation dynamics in terms of the time evolution of the thermally averaged expectation value of the cosine of the angle between the LiF molecular axis and the linearly polarized field polarization vector, resulting from an optimization of criterion j_1 (see main text for definition). (a) Using $N = 2$ individual pulses. The full line corresponds to an optimization taking into account an initial Boltzmann averaging at $T = 5$ K. The dotted line corresponds to the application of the field optimized for $T = 0$ K (see Fig. 5.1) to an initial ensemble at $T = 5$ K. (b) Using $N = 3$ pulses optimized for $T = 5$ K.

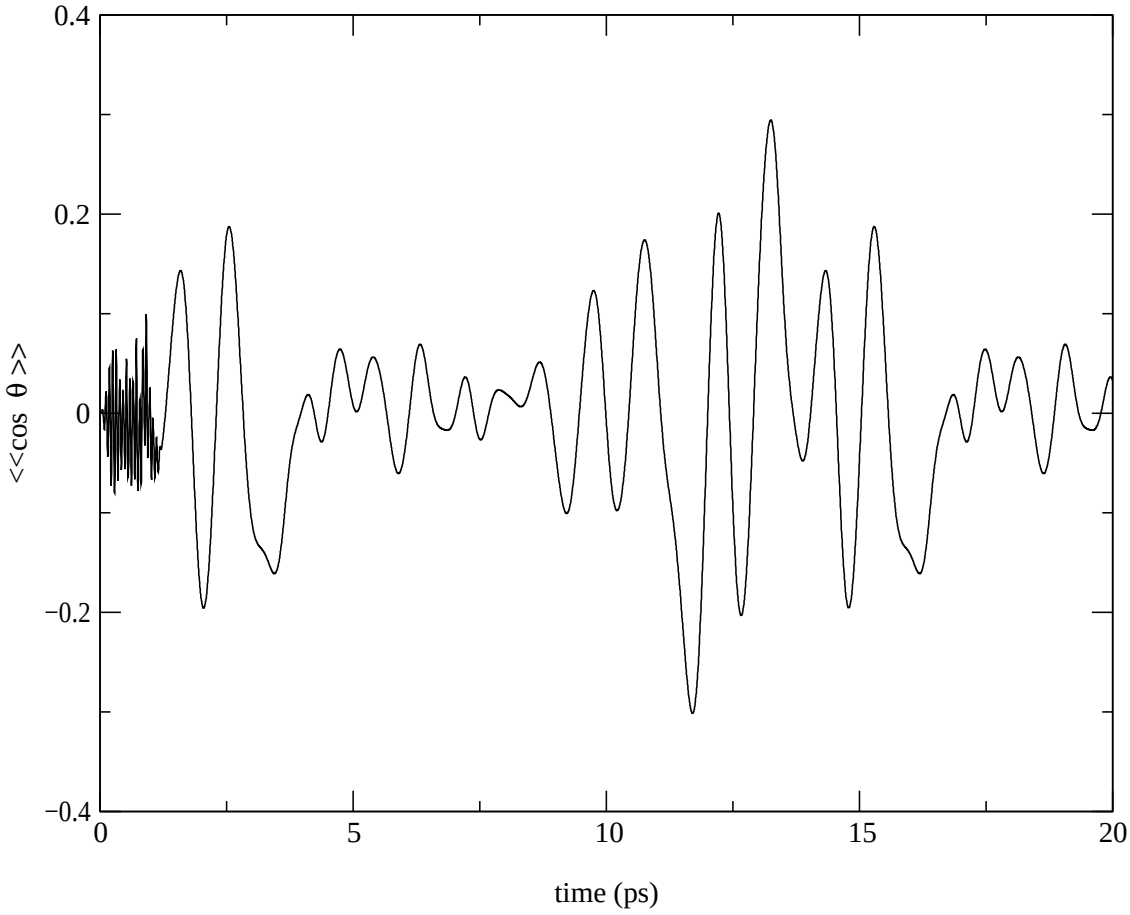


FIG. 5.6 – Same as Fig. 5.5(b), but with criteria j_4.

Chapitre 6

Optimal laser control of molecular systems : some numerical results

Ce chapitre est la reproduction d'un article qui va paraître dans *Proceedings of CDC : IEEE 2002 Conference on Decision and Control* [P5]. Dans ce chapitre on présente en plus des résultats sur le problème de l'orientation moléculaire un modèle de mécanique classique qui approxime le modèle quantique.

Optimal laser control of molecular orientation : some numerical results

A. BEN HAJ YEDDER[*†]

Abstract: We present some numerical results related to the laser control of molecular orientation. The goal is to orient a linear molecule in the direction of the linearly-polarized laser field. We either want to have the molecule oriented with the field to a high degree at (at least) one time during the time interval considered or we want this orientation to be kept as long as possible, even if it is not as good.
The control parameters are the laser field parameters : frequencies, relative phases, maximum field amplitudes and the times determining the envelope shapes. We use different objective functions measuring the orientation which are computed by solving the time-dependent Schrödinger equation including the interaction term between the laser field and the molecular system. We only consider the case of a rigid rotor, i.e., where the wave function depends only on one space variable (angle).
In order to optimize the laser field we use two different classes of algorithms : gradient-like algorithms and evolutionary algorithms (EAs).
Some promising results have been obtained and are presented here.

[*]This is a joint work with : O. ATABEK, A. AUGER, E. CANCES, C. M. DION, A. KELLER and C. LE BRIS.

[†]CERMICS, École Nationale des Ponts et Chaussées 6 & 8, avenue Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne-La-Vallée Cedex 2, FRANCE. e-mail : benhaj@cermics.enpc.fr

6.1 Introduction

The molecular system we study is the linear HCN (hydrogen cyanide) or LiF (lithium fluoride) molecule subjected to a laser field $\overrightarrow{\mathcal{E}}(t)$. Our purpose is to control

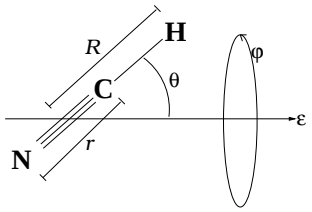


FIG. 6.1 – Model for the HCN molecule.

the *orientation* of the molecular system which is a significant step towards controlling chemical reactions. The evolution of a molecular system subjected to a laser field $\vec{\mathcal{E}}$ is modeled by the time-dependent Schrödinger equation :

$$\begin{cases} i\hbar\dfrac{\partial\psi}{\partial t} = H_0\,\psi + \vec{\mathcal{E}}(t)\cdot\vec{D}(\vec{\mathcal{E}}(t))\,\psi, \\ \psi(t=0) = \psi_0. \end{cases} \tag{6.1}$$

In this equation, the wave function ψ is assumed to depend only on the coordinates of the various nuclei the molecular system is composed of (see Figure 6.1). The presence of the electrons is accounted for through an effective potential acting on the nuclei, and contained in the Hamiltonian H_0 of the free system (when the laser is turned off). We denote by $\vec{D}(\vec{\mathcal{E}}(t))$ the dipole moment of the molecule in the presence of an external electric field $\vec{\mathcal{E}}(t)$; at the first order of perturbation theory, one can use the form $\vec{D}(\vec{\mathcal{E}}(t)) = \vec{\mu}_0 + \overline{\overline{\alpha}}\vec{\mathcal{E}}(t)$.
For this problem the control is *bilinear* (the control $\vec{\mathcal{E}}$ multiplies the state ψ) thus the mathematical theoretical results on bilinear control are very rare. For the finite dimensional approximation of equation (6.1), there exist some results that can also be extended to the infinite dimensional case. We refer to the work of G. Turinici et al. [17–19] for some recent progress on the theory of exact controllability for systems such as those we deal with here. For the optimal control problem, *some* minor things can be done. We refer in particular to [3] proving the existence of an optimal field in a very academic and simplified setting.
This paper is organized as follows : in the next Section we give more details on the problem under study. In Section 6.3 we present the different optimization methods we have used and in Section 6.4 we give and detail the results obtained for this problem.

6.2 The control problems

6.2.1 The molecular system

The Hamiltonian H in equation (6.1) can be written in a very convenient way using the so-called Jacobi coordinates ($\mathbf{R} = (R, r), \theta, \varphi$) to parameterize the state of the molecule (see Figure 6.1 for the case of the HCN molecule). As a first step toward the treatment of a more sophisticated model, we consider the case of a rigid rotor : the problem depends only on the angular variables θ, ϕ. Furthermore, symmetry conservation around the laser polarization axis allows us to separate the motion in ϕ from the motion in θ, and consider only the latter in our calculations. The Hamiltonian H therefore reduces to

$$H = H(\theta, t) = H_{rot}(\theta) + H_{laser}(\theta, t), \tag{6.2}$$

with

$$H_{rot}(\theta) = -B\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right),$$

and

$$H_{laser}(\theta, t) = -\mu_0(R, r)\mathcal{E}(t)\cos\theta - \frac{\mathcal{E}^2(t)}{2}\left[\alpha_{\parallel}(R, r)\cos^2\theta + \alpha_{\perp}(R, r)\sin^2\theta\right],$$

where R and r are fixed at their equilibrium value. In the former formulas, B is the rotational constant and μ_0 is the permanent dipole moment. The coefficients $\alpha_{\parallel}$ and $\alpha_{\perp}$ are respectively the parallel and the perpendicular components of the diagonal polarizability tensor $\overline{\overline{\alpha}}$ given by $\alpha_{\parallel} = \alpha_{zz}$ and $\alpha_{\perp} = \alpha_{xx} = \alpha_{yy}$ when (Oz) is the molecular axis.
The Schrödinger equation (6.1) depending only on the variable θ is numerically solved with a FORTRAN program written by C. M. Dion [5, 7] which uses an operator splitting method [11] coupled with a FFT for the kinetic part [4, 15].
In order to measure the orientation [12] at time t, we introduce the instantaneous criterion $j(t)$ used to compute the cost function $J(\mathcal{E})$ (see Section 6.2.4) :

$$j(t) = \langle\cos\theta\rangle = \int_0^{\pi} \cos\theta \;\; \mathcal{P}(\theta, t)\sin\theta\, d\theta, \tag{6.3}$$

where $\mathcal{P}(\theta, t)$ is the angular distribution of the molecule. In the case of the rigid rotor the angular distribution is reduced to $\mathcal{P}(\theta, t) = \|\psi\|^2_{\mathbb{C}}$ where $\|\psi\|^2_{\mathbb{C}}$ denotes the squared norm of the complex ψ. The instantaneous criterion therefore becomes

$$j(t) = \int_0^{\pi} \cos\theta \;\; \|\psi\|^2_{\mathbb{C}}\sin\theta\, d\theta. \tag{6.4}$$

The instantaneous criterion $j(t)$ takes its values in the range $[-1, 1]$, the values -1 and 1 corresponding respectively to a molecule pointing in the direction of the laser field polarization axis and in the opposite direction.

6.2.2 The classical model

From the classical mechanics viewpoint, the molecular system can be seen as a classic rotor with a permanent dipolar momentum subjected to an electric field. The rotor system is governed by the following equations [7] :

$$\begin{cases} I\ddot{\theta} + \mu_0 \mathcal{E}(t) \sin\theta + \mathcal{E}^2(t)[\alpha_\perp - \alpha_\parallel] \sin\theta\cos\theta = 0, \\ \theta(t=0) = \theta^0. \end{cases} \tag{6.5}$$

In order to approximate the quantum rotor, we consider n classical rotors governed by equation (6.5) with a $\sin\theta$ uniform distribution, of the initial condition $(\theta_k^0)_{1\leq k\leq n}$. The classical instantaneous criterion is given by

$$j_{class}(t) = \sum_{k=1}^{n} \cos(\theta_k(t)).$$

Equation (6.5) is solved using a Verlet scheme [1].

6.2.3 Choice of the laser field

The laser field $\mathcal{E}(t)$ used is the sum of N individual linearly-polarized pulses :

$$\mathcal{E}(t) = \sum_{n=1}^{N} \mathcal{E}_n(t) \sin\left(\omega_n t + \phi_n\right).$$

The envelope functions $\mathcal{E}_n(t)$ are of given sine-square form,

$$\mathcal{E}_n(t) = \begin{cases} 0 & \text{if} \quad t \leq t_{0n} \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t-t_{0n}}{t_{1n}-t_{0n}}\right)\right] & \text{if} \quad t_{0n} \leq t \leq t_{1n} \\ \mathcal{E}_{0n} & \text{if} \quad t_{1n} \leq t \leq t_{2n} \\ \mathcal{E}_{0n} \sin^2\left[\frac{\pi}{2}\left(\frac{t_{3n}-t}{t_{3n}-t_{2n}}\right)\right] & \text{if} \quad t_{2n} \leq t \leq t_{3n} \\ 0 & \text{if} \quad t \geq t_{3n} \end{cases}$$

each pulse being characterized by a set of 7 adjustable parameters, namely its frequency ω_n, relative phase ϕ_n, maximum field amplitude $\mathcal{E}_{0n}$, together with 4 times determining its shape (origin t_{0n}, rise time $t_{1n}-t_{0n}$, plateau $t_{2n}-t_{1n}$, and extinction time $t_{3n}-t_{2n}$). All beams are polarized along the same axis. We use up to 10 fields, this makes a total of $7\times 10 = 70$ parameters. In practice, optimization is made with 2 or 3 laser fields, which give a solution that we can easily analyze.

6.2.4 Choice of the cost function

Our physical goal is to orient the molecule "as best as possible"and/or "as long as possible"(compared to the rotation period of the molecule, namely 11.4 ps for HCN and 12.8 for LiF). We use different cost functions to formulate these different goals :

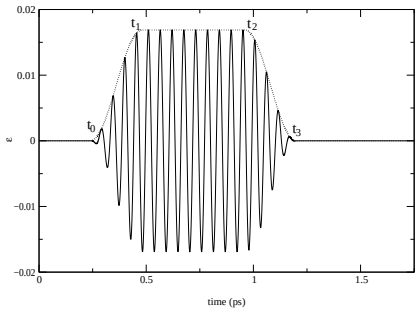


FIG. 6.2 – A typical laser field $\mathcal{E}_i(t)$.

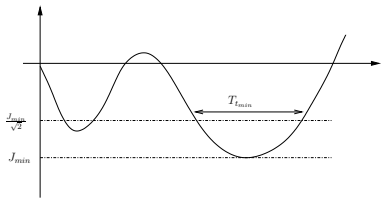


FIG. 6.3 – Construction of the hybrid criterion.

– When we want to have the molecule oriented with the field in a very good way at (at least) one time during the interval of time considered, the criterion is :

$$J = J(\mathcal{E}) = \min_{t\in[0,T]} j(t). \tag{6.6}$$

– When we want this orientation to be kept as long as possible, even if it is not so perfect, the criterion used is :

$$J = J(\mathcal{E}) = \frac{1}{T}\int_0^T j(t)\,dt. \tag{6.7}$$

– We introduce another criterion aimed at approaching the two goals together : obtaining at some given time a good orientation and keeping it as long as possible. Thus, we define a new criterion

$$J = \max(J_{min}, -J_{kept}), \tag{6.8}$$

where $J_{min} = \min_{t\in[0,T]} j(t)$ and $J_{kept} = \dfrac{T_{t_{min}}}{T}$ where $T_{t_{min}}$ is the length the connex component of $\{t \in [0,T] \mid J_{min} \leq j(t) \leq \frac{J_{min}}{\sqrt{2}}\}$ including $t_{min} = \sup\{t \mid J(t) = J_{min}\}$ (see Figure 6.3). In this criterion J_{min}, measures the way the molecule is oriented and J_{kept} measures how long the orientation is kept. The criterion ensures that J_{min} and $-J_{kept}$ are simultaneously minimized.

6.3 The optimization methods

The way we have tackled the optimization of the orientation problem is based on two different classes of algorithms : first the gradient-like algorithms (when the cost function $J(\mathcal{E})$ is differentiable, as the one defined by equation (6.7)), and second the evolutionary algorithms (EAs).
Gradient-like algorithms : We use the Polak-Ribière nonlinear conjugated gradient algorithm with a Wolfe or Goldstein-Price line-search and the BFGS algorithm. The gradient is computed using the Automatic Differentiation tool O∂yssée [10, 20] in its adjoint mode.
Evolutionary algorithms : We use two kinds of EAs : the first one is based on a classical genetic algorithm (GA) with a real representation (roulette wheel selection and *barycentric* or *multi-point* crossover) and the second one is the evolution strategies (ES) algorithm taken from EOlib [9]. Both algorithms use specific operators and some specific features, which are known to improve the performances : adaptive mutation, mutation strength decreasing with time, rescaling of the cost function...
We also tested hybrid approaches combining gradient-like algorithms and genetic algorithms. We tested GA with mutation by gradient : one parent is replaced by the result of a few iterations of a conjugated gradient algorithm using the parent as initial value.

6.4 Results

The first results we have obtained using gradient-like algorithms showed the need to use stochastic methods, since they converge after a few iterations towards a local minimum close to the initial guess (the cost function presents numerous local minima). The best results we present in this section have thus been obtained using EAs by optimizing upon a superposition of two or three lasers. Results have also been obtained by optimization with 10 lasers, but these results give lasers fields too complicated to be easily analyzed.
All results presented in this section are for the quantum model except in Section 6.4.4. In Section 6.4.1 we present the best results obtained for the different criteria presented in Section 6.2.4. One result (*kick field*) is detailed in Section 6.4.2 and the mechanism it shows is extended to the train of kick mechanism in Section 6.4.3. In Section 6.4.4 we give results for the classical mechanics approximation of the rigid rotor and in Section 6.4.5 we study the effect of temperature on orientation.

6.4.1 Results for the different criteria

Figures 6.4 and 6.5 show the optimized fields and their instantaneous criterion $j(t)$ obtained respectively with criteria (6.6) and (6.7). They have been obtained with a non-isotropic ES and with GA, respectively. However, let us emphasize that the two algorithms give similar results. Indeed, GA has given fields and instantaneous

criterion of the same form as the ones shown on Figure 6.4 and ES has also given results of the form shown on Figure 6.5.
As it may be noticed on Figure 6.4, the minimum value of $j(t)$, namely -0.46, is less than that on Figure 6.5 but the orientation does not last as long, which is expected in view of the criterion chosen. On Figure 6.6, we show a field obtained

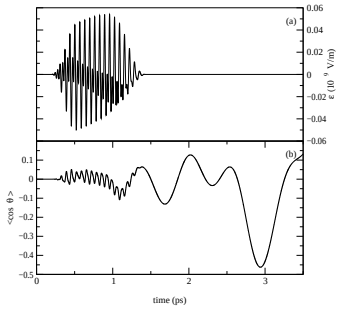


FIG. 6.4 – Best result for $J = \min_{t\in[0,T]} j(t)$. Optimization made for HCN with 2 laser fields.

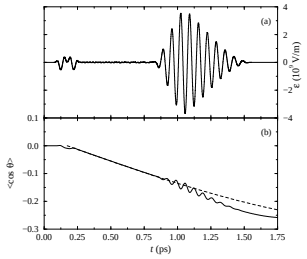


FIG. 6.5 – Best result for $J = \frac{1}{T}\int j(t)\,dt$. Optimization made for HCN with 3 laser fields.

with the hybrid criterion given by equation 6.8 and which is a succession in time of two fields with a short overlay time. We can see that as expected the orientation is maintained for a relatively long time. The main results have been obtained for the orientation problem by EAs but deterministic algorithms can be useful for a local search. We have tested how they could improve the result when used only at the end of a stochastic search. For this purpose we have first made an optimization on criterion (6.7) using GA (the result is presented on Figure 6.7 with dotted lines) and then, we have applied the conjugate gradient (CG) algorithm using the laser field obtained by GA as an initial guess for the CG algorithm. After 100 CG iterations, the criterion is improved as may be seen on Figure 6.7 with solid lines.

6.4.2 Kick mechanism

The laser field presented on Figure 6.5, called a *kick field*, is one of our most striking result from a physical viewpoint [2,8]. The kick mechanism can be analyzed

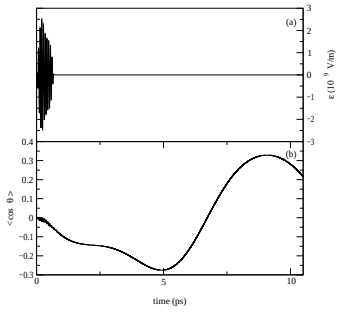


FIG. 6.6 – Best result for the hybrid criterion given by equation 6.8. Optimization made for HCN with 2 laser fields.

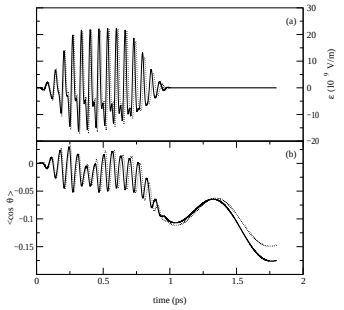


FIG. 6.7 – Optimization by CG after optimization by GA.

using the "sudden-impact" model [6] as the pulse duration is much shorter that the rotational period. Comparison with the results from a sudden-impact approximation [6] was made for this laser field (see Figure 6.8) and also for a constructed kick field. This field was also tested with the LiF molecule and we observe the same orientation mechanism.

6.4.3 Train of kicks

Another idea consists in starting with a field previously classified as a kick shape and using a succession of such fields in order to orient the molecule. The purpose of the optimization is thus to find the good delay between two successive *kicks*. Indeed we hope that by kicking several times the molecule we can lower the instantaneous criterion.
Figure 6.9 (a), is the result of an optimization of the criterion (6.7) with ES and it clearly illustrates the idea of kicking several times the molecule. We emphasize that for this result the instantaneous criterion remains for a long time under the value -0.2. Figure 6.9 (b) is the result of the optimization with the criterion (6.6). The criterion value (-0.82) is the best value we have ever had. However, the production of such fields remains an experimental challenge.

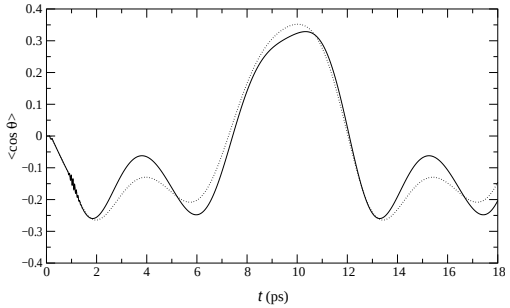


FIG. 6.8 – (a) : orientation value $\langle \cos\theta \rangle$ for HCN molecule with the kick field (solid line) and with the sudden-impact model (dashed line).

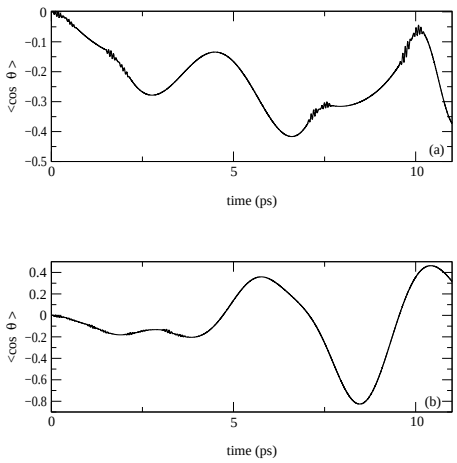


FIG. 6.9 – Optimization with train of kicks for HCN molecule. (a) : Best result for $J = \frac{1}{T}\int j(t)\,dt$. (b) : Best result for $J = \min_{t\in[0,T]} j(t)\,dt$.

6.4.4 Results for the classical model

On Figure 6.10 we can see that the classical rotor is oriented in a similar way as the quantum rotor when the kick field is applied. The difference of orientation degree is due to the quantum effects which can not be modelized by classical mechanics. We

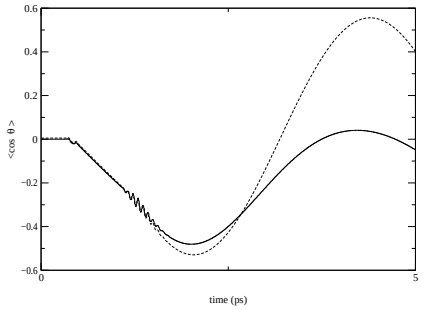


FIG. 6.10 – Orientation value $\langle \cos\theta \rangle$ for LiF molecule with the kick field for the quantum model (solid line) and for the classical model (dashed line).

also made optimization with the classic rotor and then tested the optimized laser field obtained on the quantum rotor. On Figure 6.11 we can see that the orientation degree is less important for the quantum rotor than for classic rotor again due to quantum effects.

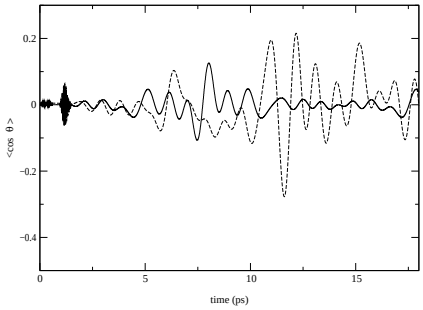


FIG. 6.11 – Orientation value $\langle \cos\theta \rangle$ for LiF molecule with an optimized filed for classical model (dashed line) and with the same field for the quantum model (solid line).

6.4.5 Temperature effects

Some previous studies in the literature have shown the fast decrease of the degree of alignment or orientation with increasing temperature [13,14,16] (see Figure 6.12). With a temperature $T > 0$, the initial state distribution of the spheric harmonics Y_l^m is given by the partition function

$$Q = \sum_J (2J+1) \exp\left[\frac{-BJ(J+1)}{k_B T}\right],$$

where B is the rotational constant and k_B is the Boltzmann constant. The thermal averaged orientation measure is given by

$$\begin{aligned} \langle\langle \cos\theta \rangle\rangle (t) &= Q^{-1} \sum_{J} \exp\left[\frac{-BJ(J+1)}{k_B T}\right] \\ &\times \sum_{M=-J}^{J} \langle \cos\theta \rangle_{J,M}(t). \end{aligned} \tag{6.9}$$

An optimized laser field for the initial state $J = 0$, $M = 0$ is no more appropriate for orienting an excited initial state. With a temperature $T = 5K$, the dominant initial states are $J = 1$, $M = -1, 0, 1$. An optimized laser field for the initial state $J = 1$, $M = 0$ (we know that the initial states $J = 1$, $M = -1$ and $J = 1$, $M = 1$ cannot be oriented because of selection rules) loses also in its orientation degree when applied to the thermal distribution of states at $T = 5K$.
In order to obtain orientation for the excited initial state we run an optimization using $j_{avr}(t) = \langle\langle \cos\theta \rangle\rangle (t)$ as instantaneous criterion. The first results (see Figure 6.13) show that optimization of the cost function 6.6 with $j_{avr}(t)$ is possible and perform better the orientation of the excited state. We are currently working on this optimization in order to improve the orientation and to test the different cost functions.

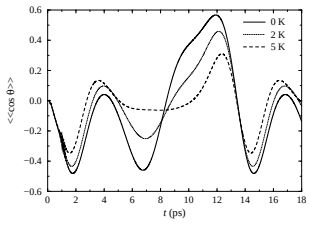


FIG. 6.12 – Temperature effect : averaged orientation value $\langle\langle \cos\theta \rangle\rangle$ for LiF molecule submitted to the kick filed.

6.5 Conclusion

In this paper we have presented an optimal control approach for the molecular orientation by laser fields. Some interesting and promising results have already been obtained by this approach. The same approach was used and easily adapted to different models : classic/quantum model and thermal averaged orientation model. Other approaches, in particular based on better objective functions, are currently under study.

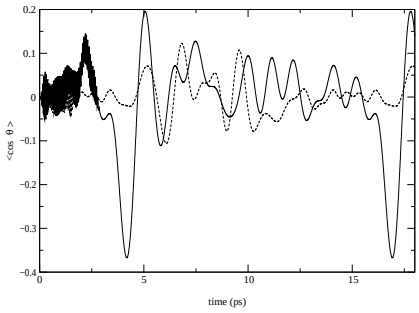


FIG. 6.13 – Temperature effect : averaged orientation value $\langle\langle \cos\theta \rangle\rangle$ for LiF molecule at $T = 5K$ with an optimized field for $J = 1$, $M = 0$ (dashed line) and with an optimized field for the thermal distibution of states (solid line).

6.6 Acknowledgements

This work is financially supported by the *Action Concertée Incitative Jeunes Chercheurs* from the French Ministry of Research and Technology.

References

[1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, 1996.

[2] A. Auger, C. M. Dion, A. Ben Haj Yedder, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Optimal laser control of molecular systems : methodology and results. M3AS : to appear.

[3] E. Cancès, C. Le Bris, and M. Pilot. Contrôle optimal bilinéaire sur une équation de Schrödinger. *Note aux Compte Rendu de l'Académie des Sciences, Série 1*, 330 :567–571, 2000.

[4] C. E. Dateo and H. Metiu. Numerical solution of the time-dependent schrodinger equation in sperical coordinates by fourier-transformation methods. *J. Chem. Phys*, 95 :7392–7400, 1991.

[5] C. M. Dion. *Dynamique de l'alignement et de l'orientation moleculaire induite par laser. Simulations numeriques sur HCN en champ infrarouge.* PhD thesis, Université de Sherbrooke et Université de Paris-Sud, 1999.

[6] C. M. Dion, A. Keller, and O. Atabek. Orienting molecules using half-cycle pulses. *Eur. Phys. J. D*, 14 :249–255, 2001.

[7] C. M. Dion, A. Keller, O. Atabek, and A. D. Bandrauk. Laser-induced alignment dynamics of HCN : Roles of the permanent dipole moment and the polarizability. *Phys. Rev. A*, 59 :1382–1391, 1999.

[8] C. M. Dion, A. Ben Haj Yedder, E. Cancès, A. Keller, C. Le Bris, and O. Atabek. Optimal laser control of orientation : The kicked molecule. submitted to Phys. Rev. A.

[9] EO. C++ class library, http ://eodev.sourceforge.net/.

[10] C. Faure and Y. Papegay. O∂yssée User's Guide Version 1.7. *Rapport Technique INRIA RT-0224*, 1998.

[11] M.D. Feit, J.A. Fleck, and A. Steiger. Solution of a Schrodinger equation by a spectral method. *J. Comput. Phys.*, 47 :412–433, 1982.

[12] B. Freidrich and D. R. Herschbach. On the possibility of orienting rotationally cooled polar molecules in an electric field. *Z. Phys. D*, 18 :153–161, 1991.

[13] Mette Machholm. Postpulse alignment of molecules robust to thermal averaging. *J. Chem. Phys.*, 115 :10724–10730, 2001.

[14] Mette Machholm and Niels E. Henriksen. Field-free orientation of molecules. *Phys. Rev. Lett.*, 87 :193001, 2001.

[15] R. Numico, A. Keller, and O. Atabek. Laser-induced molecular alignment in dissociation dynamics. *Phys. Rev. A*, 52 :1298–1309, 1995.

[16] Juan Ortigoso, Mirta Rodríguez, Manish Gupta, and Bretislav Friedrich. Time evolution of pendular states created by the interaction of molecular polarizability with a pulsed nonresonant laser field. *J. Chem. Phys.*, 110 :3870–3875, 1999.

[17] G. Turinici. Controlabilité exacte de la population des états propres dans les systèmes quantiques bilinéaires. *Note aux Comptes Rendus de l'Académie des Sciences, Série 1*, 330 :327–332, 2000.

[18] G. Turinici and H. Rabitz. Quantum wave function controllability. *Chem. Phys.*, 267 :1–9, 2001.

[19] G. Turinici and H. Rabitz. Wavefunction controllability in quantum systems. *Preprint*, 2001.

[20] A. Ben Haj Yedder, E. Cancès, and C. Le Bris. Optimal laser control of chemical reactions using automatic differentiation. In George Corliss, Christèle Faure, Andreas Griewank, Laurent Hascoët, and Uwe Naumann (eds.), editors, *Proceedings of Automatic Differentiation 2000 : From Simulation to Optimization*, pages 203–213, New York, 2001. Springer-Verlag.

Chapitre 7

Optimisation de la génération d'harmoniques hautes (HHG) pour la création d'un laser attoseconde

Ce chapitre présente une étude encours menée avec S. Chelkowski et O. Atabek. Le but de notre étude est de créer un champ laser attoseconde, champ très court dont la durée est de l'ordre de quelques $10^{-18}s$ en utilisant la méthode de glissement de fréquence (chirp). On présente ici les premiers résultats obtenus dans le cadre de cette étude.

7.1 Introduction

Un des buts des générations d'harmoniques hautes (HHG : High Harmonic Generation) est de convertir des lasers standards (UV) en lasers de hautes fréquences (rayon X). La génération d'harmoniques hautes peut aussi permettre d'obtenir des impulsions lasers ultra courtes, d'une durée le de moins de $10^{-15}s$. Pour cela , on excite à l'aide d'un laser intense un atome ou une molécule. Ceux-ci émettent alors des photons d'une fréquence multiple (impaire) de la fréquence du laser initial. L'intensité des photons émis décroît avec leur énergie. Ce sont ces photons émis que l'on utilise pour créer le nouveau champ laser. Les caractéristiques de ce nouveau champ laser sont déterminées par le choix des photons sélectionnés en fonction de leur fréquence, et donc leur énergie ($E = \hbar\nu$), et de leur intensité.

Dans certaines études théoriques et expérimentales [3, 5, 9] les auteurs se sont intéressés à une fréquence donnée qu'ils ont favorisé par rapport aux autres fréquences. Ceci peut avoir comme application la création d'un champ laser d'une certaine fréquence qui sera intense et d'amplitude continue. Le but de notre étude est de favoriser la création d'un champ laser *attoseconde*, champ très court dont la durée est de l'ordre de quelques $10^{-18}s$. A l'heure actuelle l'objectif (dans cette étude et expérimentalement) est d'atteindre une durée de quelques centaines d'attosecondes. Dans des études précédentes des trains de pulses attoseconde sont générés [1, 9] et notre objectif est d'isoler de ce train un pulse en augmentant son intensité par rapport aux autres pulses. Pour cela on utilise un champ laser intense $800nm$, d'une intensité de l'ordre de $2 \times 10^{14} W/cm^2$ et d'une durée de l'ordre de $10fs$. Les paramètres de ce champ laser sont optimisés par un algorithme génétique.

7.2 Le modèle physique

Dans ce problème on considère un atome d'hydrogène soumis à un champ laser $\mathcal{E}(t)$ polarisé linéairement [12]. Nous considérons un modèle uni-dimentionnel, où le noyau est supposé fixe et l'électron se déplace selon un axe (Oz) défini par l'axe de polarisation du laser [8]. L'équation de Scrhödinger dépendante du temps de ce modèle s'écrit sous la forme :

$$\begin{cases} \dfrac{\partial}{\partial t}\psi(z,t) = H\psi(z,t), \\ \psi(z,t=0) = \psi^0(z), \end{cases} \tag{7.1}$$

où l'hamiltonien H est donné par

$$H = -\frac{1}{2}\frac{\partial^2}{\partial z^2} + V(z,t), \tag{7.2}$$

avec

$$V(z,t) \;=\; \frac{-1}{\sqrt{z^2+1}} \;+ z\mathcal{E}(t). \tag{7.3}$$

Cette équation est résolue par une méthode de décomposition d'opérateurs (operator splitting) [2, 6, 7] couplée une transformée de Fourier rapide (FFT) pour la partie cinétique.

La solution $\psi(z,t)$ de l'équation (7.1) permet de calculer le spectre $\hat{\mathcal{E}}^a(\omega)$ des photons émis par l'atome d'hydrogène sous l'effet de l'excitation du champ laser [4, 13]. Ce spectre est donné par :

$$\hat{\mathcal{E}}^a(\omega) = \int_{-\infty}^{+\infty} \ddot{d}(t) e^{-i\omega t} dt, \tag{7.4}$$

où l'accélération $\ddot{d}(t) = \dfrac{d^2}{dt^2} d(t)$ avec :

$$d(t) = \langle \psi(z,t)|z|\psi(z,t)\rangle_z = \int_{-\infty}^{+\infty} z|\psi(z,t)|^2 dz. \tag{7.5}$$

L'accélération $\ddot{d}(t)$ est calculée en utilisant l'égalité suivante [4] :

$$\ddot{d}(t) = \int_{-\infty}^{+\infty} \left[-\frac{d}{dz} V(z,t)\right] |\psi(t,z)|^2 dz.$$

Le pulse attoseconde correspondant à une harmonique donnée de fréquence ω_c est calculé par une transformée de Fourier du spectre $\hat{\mathcal{E}}^a(\omega)$ auquel on a ajouté un filtre :

$$\mathcal{E}^a(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f_{\omega_c}(\omega) \hat{\mathcal{E}}^a(\omega) e^{i\omega t} dw, \tag{7.6}$$

où $f_{\omega_c}(\omega)$ est un filtre gaussien centré en ω_c. Cette fréquence centrale ω_c est soit fixée dès le départ à une valeur correspondante à une harmonique haute (dans la zone entre la $20^{\text{ème}}$ et la $30^{\text{ème}}$ harmonique par exemple) soit laissée libre et trouvée par l'algorithme d'optimisation. Le rôle de ce filtre, qui est utilisé expérimentalement, est de couper les fréquences des harmoniques non désirées.

Pour effectuer les calculs présentés dans cette section on utilise un programme écrit en Fortran et fourni par S. Chelkowski.

7.3 Les paramètres de contrôle

On utilise dans ce problème la méthode de *glissement de fréquence* (chirp) [10]. Cette méthode consiste à utiliser un champ laser d'une fréquence ω_0 donnée que l'on fait varier au cours du temps. Le champ laser utilisé s'écrit sous la forme :

$$\mathcal{E}(t) = \mathcal{E}_0(t) \cos\left[\phi(t)\right]$$

où l'enveloppe $\mathcal{E}_0(t)$ est donnée par :

$$\mathcal{E}_0(t) = E_0 \sin^2\left(\frac{\pi t}{T}\right),$$

et où $\phi(t)$ est donnée par :

$$\phi(t) = \omega_0 t + \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4.$$

La fréquence instantanée est donnée par :

$$\omega(t) = \frac{d}{dt}\phi(t) = \omega_0 + \alpha_1 + 2\alpha_2 t + 3\alpha_3 t^2 + 4\alpha_4 t^3.$$

Dans les équations précédentes, E_0 est l'intensité du laser, T représente le temps total et la quantité $(\omega(t) - \omega_0)$ représente le glissement de fréquence qui ne doit pas dépasser en valeur absolue $5\%\omega_0$ sur l'intervalle $[0, T]$.

Les paramètres de contrôle sont l'intensité E_0, les paramètres de glissement de fréquence α_0, α_1, α_2, α_3, α_4 et éventuellement, le centre du filtre ω_c définie dans la section précédente.

7.4 Les critères optimisés

La boucle d'optimisation a été réalisée par l'algorithme génétique développé pour le problème de contrôle par laser de l'orientation moléculaire [11]. Dans la boucle d'optimisation on a utilisé successivement deux critères. Le premier critère (critère indirect), consiste à s'intéresser au spectre émis par l'atome d'hydrogène $\hat{\mathcal{E}}^a(\omega)$ et de choisir quelques harmoniques voisines que l'on optimise dans un certain sens dans le but d'améliorer la forme du pulse attoseconde généré $\mathcal{E}^a(t)$. Le second critère (critère direct), a été construit directement à partir du pulse attoseconde généré $\mathcal{E}^a(t)$. En effet, les résultats obtenus après l'optimisation du premier critère on montré que ce dernier est insuffisant car le pulse $\mathcal{E}^a(t)$ généré donne une impulsion assez courte mais qui n'est pas séparée des autres impulsions plus faibles qui l'entourent (voir figure 7.4). La figure 7.1 illustre les schémas d'optimisation utilisés dans ces deux cas.

7.4.1 Critère indirect

Dans ce critère on se donne une harmonique centrale H_{j_0} du spectre $\hat{\mathcal{E}}^a(\omega)$ émis par l'atome d'hydrogène et un nombre $2j_{max}$ d'harmoniques voisines. Dans les exemples traités j_0 a été choisi dans l'intervalle $[19, 29]$ et le nombre j_{max} a été pris égal à 2 ou 4. Le but est de rendre les intensités et les phases de ces harmoniques très proches afin de créer une interférence constructive qui maximise le

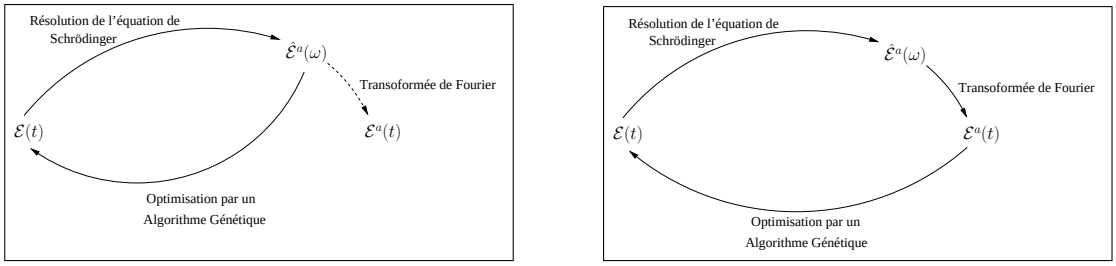


FIG. 7.1 – Illustration des schémas d'optimisations : (a) le critère J_1 est construit à partir du spectre $\hat{\mathcal{E}}^a(\omega)$, (b) le critère est construit directement à partir du pulse attoseconde $\mathcal{E}^a(t)$. Le champ $\mathcal{E}(t)$ est le champ laser utilisé qui représente le contrôle.

pulse attoseconde $\mathcal{E}^a(t)$. Le critère à minimiser qui a été utilisé s'écrit sous la forme suivante :

$$J_1 = \max_{j\in[j_0-j_{max},j_0+jmax]} \left[\left| \frac{\max(|\hat{\mathcal{E}}^a(j\omega_0)|, |\hat{\mathcal{E}}^a(j_0\omega_0)|)}{\min(|\hat{\mathcal{E}}^a(j\omega_0)|, |\hat{\mathcal{E}}^a(j_0\omega_0)|)} - 1 \right| + \left| arg(\hat{\mathcal{E}}^a(j\omega_0)) - arg(\hat{\mathcal{E}}^a(j_0\omega_0)) \right| \right],$$

où $|\hat{\mathcal{E}}^a(j\omega_0)|$ et $arg(\hat{\mathcal{E}}^a(j\omega_0))$ sont respectivement le module et l'argument du nombre complexe $\hat{\mathcal{E}}^a(j\omega_0)$ représentant l'harmonique H_j dans le spectre $\hat{\mathcal{E}}^a$. Dans ce cas le paramètre ω_c définissant le centre du filtre (voir section 7.2) est fixé à la valeur $\omega_c = j_0\omega_0$.

7.4.2 Critère direct

Ce critère est construit directement à partir du pulse attoseconde $\mathcal{E}^a(t)$ de la manière suivante :

$$J_2 = \frac{I_1}{I_2},$$

avec

$$I_1 = \frac{1}{2t_{pulse}} \int_{t_c-t_{pulse}}^{t_c+t_{pulse}} |\mathcal{E}^a(t)|dt,$$

et

$$I_2 = \frac{1}{T-2t_{pulse}} \left[\int_0^{t_c-t_{pulse}} |\mathcal{E}^a(t)|dt + \int_{t_c+t_{pulse}}^{T} |\mathcal{E}^a(t)|dt, \right],$$

où T est la durée totale de l'expérience, t_{pulse} est la durée d'une impulsion qui peut être considérée comme brève (t_{pulse} est de l'ordre de 10 u.a.) et t_c est donné par $t_c = argmax(\mathcal{E}^a(t))$. La maximisation du critère J_2 tend à maximiser l'impulsion la plus importante du pulse $\mathcal{E}^a(t)$ et à minimiser les autres impulsions autour de cette dernière. Dans cette partie le paramètre ω_c définissant le centre du filtre est l'un des paramètres de contrôle et peut varier dans l'intervalle $[10, 31]$.

7.5 Résultats

Le premier résultat a été obtenu en optimisant le critère indirect J_1 pour la $27^{\text{ème}}$ harmonique H_{27} et ses 4 harmoniques voisines $\{H_j\}_{j=23,25,29,31}$. Dans cette optimisation le centre du filtre a été fixé au niveau de l'harmonique H_{27} (on fixe $\omega_c = 27\omega_0$). Le résultat de l'optimisation a produit un champ laser (voir figure 7.2 (b)) permettant de rapprocher les intensités et les phases des 5 harmoniques considérées comme le montre la figure 7.3. La figure 7.4 montre que le pulse attoseconde obtenu dans ce cas est plus intense que le pulse donné par le champ laser non optimisé mais que les impulsions de ce pulse ne sont pas pas bien séparées. En effet, l'impulsion la plus importante (entre 10 fs et 11.3 fs sur la figure 7.4 (b)) n'est pas bien séparée des autres impulsions qui sont assez nombreuses et assez intenses. L'objectif de la génération d'un pulse attoseconde est de bien isoler l'impulsion la plus importante et de réduire les autres impulsions qui l'entourent. L'observation de ce résultat ainsi que celle d'autres résultats obtenus en considérant d'autres harmoniques a conduit à la construction du critère J_2 qui traduit mieux l'objectif de la génération de pulses attoseconde.

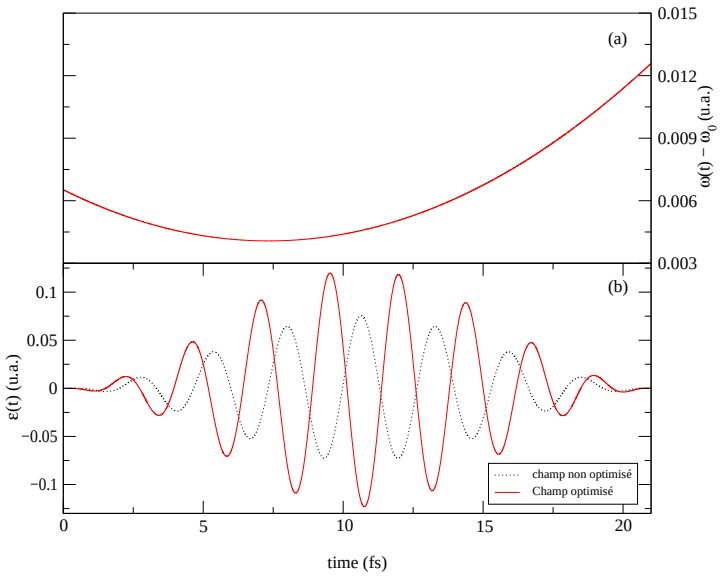


FIG. 7.2 – Résultat de l'optimisation du critère J_1. (a) : variation du glissement de fréquence en fonction du temps. (b) : champ laser excitant l'atome avec (ligne continue) et sans (ligne en pointillés) glissement de fréquence.

L'optimisation du critère J_2 a été réalisée en ajoutant le paramètre ω_c définissant le centre du filtre à l'ensemble des paramètres de contrôle. L'algorithme d'optimisation peut faire varier ce paramètre dans l'intervalle $[11\omega_0, 31\omega_0]$. Ce choix permet d'avoir un pulse attoseconde dans la gamme des lasers hautes fréquences. La valeur

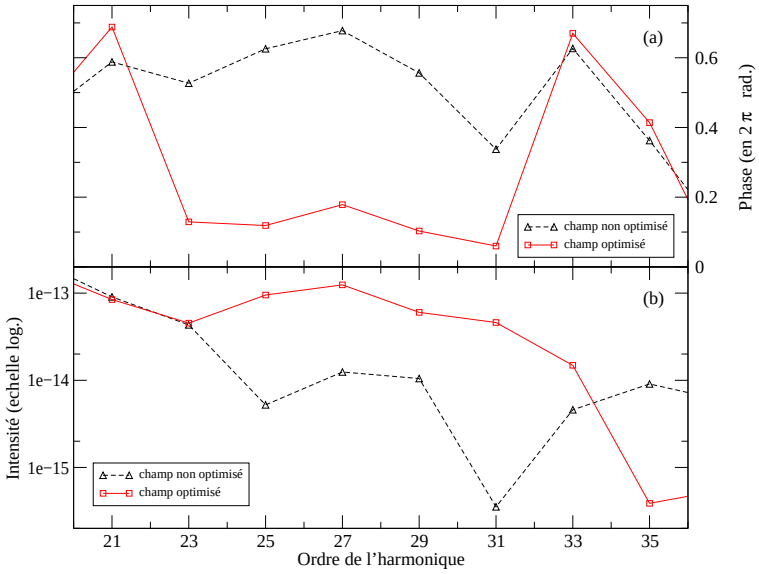


FIG. 7.3 – Résultat de l'optimisation du critère J_1 minimisant l'écart entre les phases ($arg(\hat{\mathcal{E}}^a(j\omega_0))$) les intensités ($|\hat{\mathcal{E}}^a(j\omega_0)|$) des harmoniques 23, 25, 27, 29 et 31.

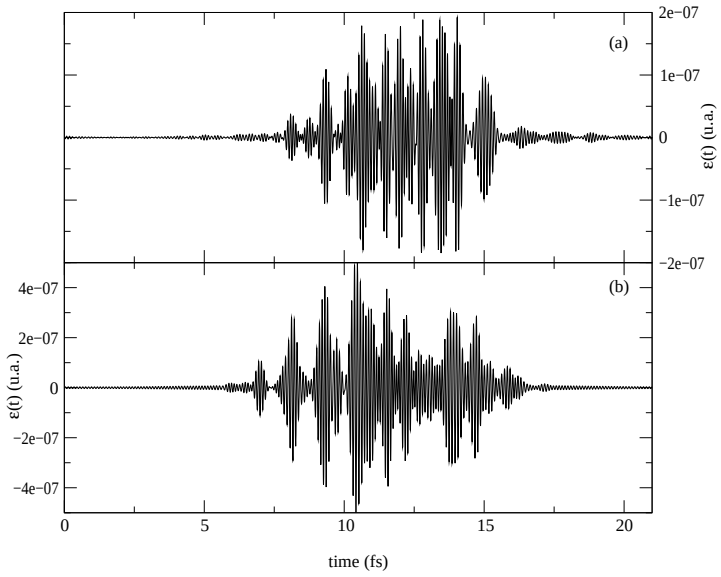


FIG. 7.4 – Le pulse attoseconde obtenu après l'optimisation (b) du critère J_1 comparé au pulse obtenu par un champ non optimisé (a). Dans les deux cas le filtre est centré au niveau de l'harmonique H_{27}.

de ω_c trouvée pour le résultat, présentée sur les figures 7.5, 7.6 et 7.7, est $\omega_c = 31\omega_0$. Les figures 7.7 et 7.4 montrent que la qualité du pulse attoseconde obtenu par l'optimisation du critère J_2 (figure 7.7 (b)) est nettement meilleure que les pulses obtenus par le champ laser non optimisé (figure 7.7 (a) et 7.4 (a)) ou par le champ laser obtenu en optimisant le critère J_1 (figure 7.4 (b)). On remarque que l'optimisation du critère J_2 ne semble rapprocher ni les intensités ni les phases des harmoniques voisines de l'harmonique H_{27} comme on peut le voir sur la figure 7.6. Dans la suite de cette étude on abandonnera le critère J_1 qui est moins bien adapté à l'objectif de la génération de pulses attoseconde.

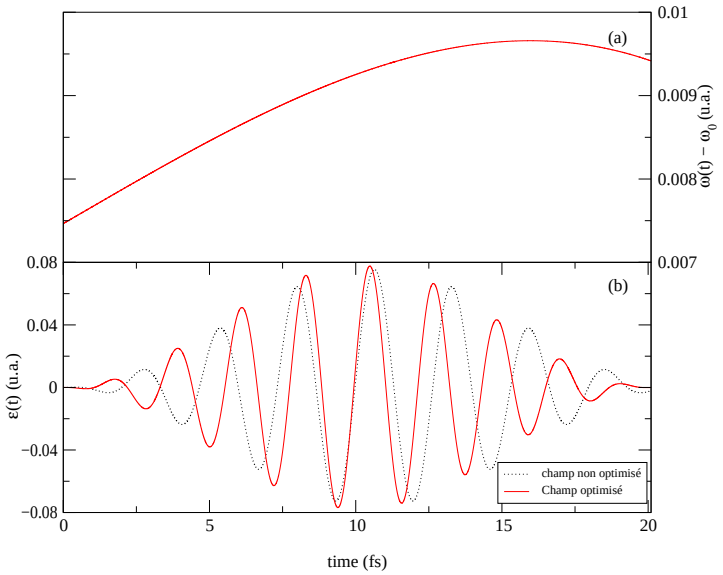


FIG. 7.5 – Résultat de l'optimisation du critère J_2. (a) : variation du glissement de fréquence en fonction du temps. (b) : champ laser excitant l'atome avec (ligne continue) et sans (ligne en pointillés) glissement de fréquence.

7.6 Conclusion

Les premiers résultats présentés dans cette étude sont encourageants et montrent que la méthode de glissement de fréquence peut permettre la création de lasers attoseconde. Cette étude est encore en cours. Actuellement on cherche à améliorer le résultat présenté sur la figure 7.7 (b) en réduisant les impulsions latérales et en augmentant l'intensité de l'impulsion centrale.

Dans la suite les résultats trouvés seront testés dans le cas tri-dimentionnel pour vérifier si la qualité des pulses attoseconde sera conservée en passant du cas 1D au cas 3D. Dans le cas où cette qualité serait perdue, une optimisation d'un modèle

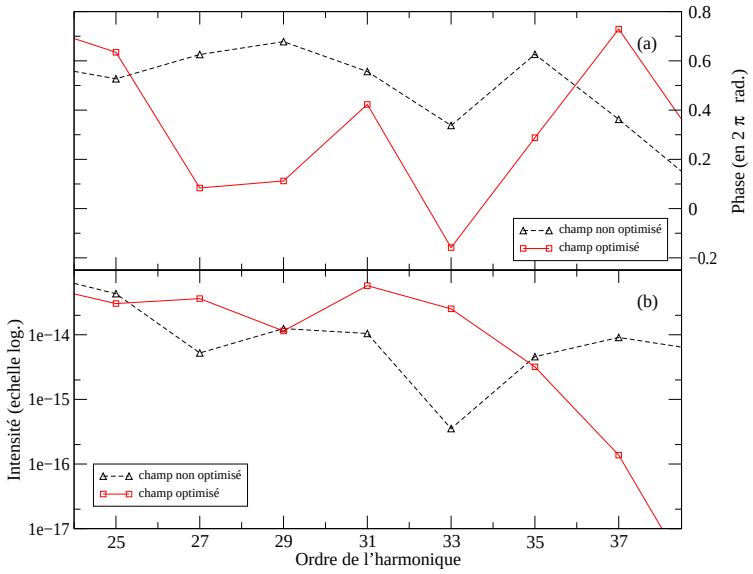


FIG. 7.6 – Résultat de l'optimisation du critère J_2. Les intensités et les phases des harmoniques voisines de l'harmonique H_{31} ne sont pas très proches.

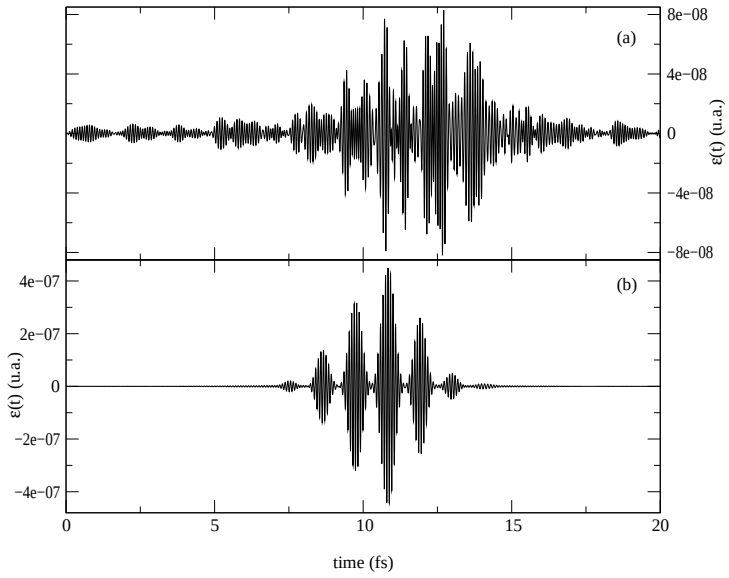


FIG. 7.7 – Le pulse attoseconde obtenu après l'optimisation (b) du critère J_2 comparé au pulse obtenu par un champ non optimisé (a). Dans les deux cas le filtre est centré au niveau de l'harmonique H_{31}. Cette harmonique est celle trouvée par l'algorithme d'optimisation.

3D est envisageable. Mais dans ce cas, il faudra d'abord améliorer le code 3D pour l'accélérer car dans sa version actuelle un calcul simple prends 36 heures (contre 5 minutes pour le code 1D) !

Une autre voie pour poursuivre ce travail est d'étudier l'effet de l'atome d'hydrogène (ou plus exactement des atomes d'hydrogène) sur le champ laser. Ceci conduit à coupler un ensemble d'équations de de Scrhödinger (du même type que l'équation 7.1) avec une équation des ondes avec un terme de couplage supplémentaire.

Références

[1] P. Antoine, A. L. Huillier, and M. Lewenstein. Attosecond pulse trains using high-order harmonics. *Phys. Rev. Lett*, 77 :1234, 1996.

[2] J. N. Bardsley, A. Szöke, and M. Comella. Multiphoton ionization from a short range potential by short-pulse lasers. *J. Phys. B*, 21 :3899–3916, 1988.

[3] R. Bartels, S. Backus, I. Christov, H. Kapteyn, and M. Murnane. Attosecond time-scale feedback control of coherent X-ray generation. *Chem. Phys.*, 267(1-3) :277–289, 2001.

[4] K. Burnett, V. C. Reed, J. Cooper, and P. L. Knight. Calculation of the background emitted during high-harmonic generation. *Phys. Rev., A.*, 45(5) :3347–3349, 1992.

[5] X. Chu and S. I. Chu. Optimization of high-order harmonic generation by genetic algorithm and wavelet time-frequency analysis of quantum dipole emission. *Phys. Rev., A.*, 64(2), 2001.

[6] M. D. Feit and J. A. Fleck. Wave packet dynamics and chaos in the hénon-heiles system. *J. Chem. Phys.*, 80 :2578–2584, 1984.

[7] M. D. Feit, J. A. Fleck, and A. Steiger. Solution of a Schrodinger equation by a spectral method. *J. Comput. Phys.*, 47 :412–433, 1982.

[8] J. Javanainen, J. H. Eberly, and Q. Su. Numerical simulations of multiphoton ionization and above threshold electron spectra. *Phys. Rev., A.*, 38 :3430, 1988.

[9] L. Roos, M. B., Gaarde, and A. L. Huillier. Tailoring harmonic generation to different applications using a genetic algorithm. *J. Phys. B*, 34 :5041, 2001.

[10] P. Salières, P. Antoine, A. de Bohan, and M. Lewenstein. Temporal and spectral tailoring of high-order harmonics. *Phys. Rev. Lett*, 81 :5544, 1998.

[11] A. Ben Haj Yedder. MyGa : a Genetic Algorithm in Fortran., http ://cermics.enpc.fr/∼benhaj/MyGa/.

[12] T. Zuo, A. D. Bandrauk, M. Ivanov, and P. B. Corkum. Control of high-order harmonic generation in strong laser fields. *Phys. Rev., A.*, 51(5) :3991–3998, 1995.

[13] T. Zuo, S. Chelkowski, and A. D. Bandrauk. Harmonic generation by the H2+ molecular ion in intense laser fields. *Phys. Rev., A.*, 48(5A) :3837–3844, 1993.

Chapitre 8

A numerical investigation of the 2-dimensional crystal problem

Ce chapitre présente des méthodes numériques développées pour traiter des problèmes d'optimisation de géométrie. Ces méthodes sont basées sur des approches déterministes (algorithme de gradient conjugué et algorithme de BFGS) et stochastiques (algorithmes génétiques). Dans l'un des cas (interaction par le potentiel de Lennard-Jones) ces méthodes ont été adaptées à ce cas et ont utilisé en particulier des résultats théoriques présentés également dans ce chapitre.

A numerical investigation of the 2-dimensional crystal problem

A. Ben Haj Yedder[1], X. Blanc[2] & C. Le Bris[1]
[1] CERMICS, École Nationale des Ponts et Chaussées,
6 & 8, avenue Blaise Pascal, Cité Descartes,
Champs sur Marne, 77455 Marne-La-Vallée Cedex 2, FRANCE
{benhaj,lebris}@cermics.enpc.fr

[2] Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
Boîte courrier 187
75252 Paris Cedex 05, FRANCE
blanc@ann.jussieu.fr

Abstract: This paper presents a numerical approach for the problem of determining whether some mathematical models for matter at zero temperature spontaneously give rise to a periodic structure. For different types of interactions, of either quantum or classical nature, we numerically compute the ground state of a set of N identical atoms, with N large. For this purpose, various optimization algorithms, both of deterministic and stochastic types, are developped and adapted. Whatever the model and the algorithm, results show that the ground state approaches a periodic structure as N grows.

8.1 Introduction

It is a long-standing open problem to understand why matter is crystalline at low temperature (see for instance [24] for a review on the topic). The problem may take many mathematical aspects, and we will here focus on one aspect amongst others. Given a molecular model, which to the position $\{X_i\}$ of a set of N atoms associates an energy $E(\{X_i\})$, the question under consideration is to prove that the ground state configuration $\{X_i^0\}_{1\leq i\leq N}$, which is the minimizer (when it exists) of

$$I_N = \inf\left\{ E(\{X_i\}), \quad X_i \in \mathbf{R}^d \right\},$$

is periodic, or at least approaches some periodic configuration as N goes to infinity. Here, $d \in \{1, 2, 3\}$ is the dimension of the ambient space.

There exist many theoretical results on the one-dimensional case ($d = 1$). The case of molecular models with two-body interaction potential is addressed in [10, 19, 20, 22, 25, 27], for a wide variety of potentials. Some cases of quantum model (in which the electrons are treated through a quantum model, and the nuclei are treated classically), have been studied in [4], still in dimension one.

On the contrary, there are very few results in dimension larger than or equal to two. All the works we know are devoted to attractive two-body potentials for hard spheres [12, 23]. A rigorous general theoretical analysis of the 2 dimensional case seems to be out of reach, even if we hope to be able to make progress in this direction in a near future. As far as the 3 dimensional case is concerned, we are not aware of any theoretical results.

On the numerical side, it is unfortunately also out of reach to determine the list of global minimizers and check whether they are periodic, even for a finite but large enough number of particles (in theory, one should have to handle an infinite number of particles as the result cannot be expected to already hold for a finite number of particles, but only in the limit of an infinite number). What we show however here is that, using a stochastic type algorithm that enjoys good exploration properties, and starting from initial guesses that are as general as possible practically, we converge to the same configuration, that seems to approach a periodic configuration as the number of particles grows. We consider this as a convincing argument tending to prove the periodicity of the global minimizer in the limit N goes to infinity.

Let us emphasize that our main concern here is indeed the behavior as N becomes large, with a view to check whether or not numerics corroborates the theoretically expected periodic behavior. This is why we use crude initial guesses and handle a large variational space. All the constraints we impose during the minimization procedure are rigorously proven. In this respect, our standpoint and aim are thus deliberately different from those of works trying to determine explicit optimized configurations of large clusters, of precise chemical constitution : see for instance [9, 11, 13, 16, 17, 28] and the references therein. Consequently, our optimization techniques also differ, and are significantly differently implemented. For instance, computational cost is definitely not our priority. We therefore do not impose extra restrictions that speed up the calculations but are not rigourously justified. In the same vein, we do not choose as initial guesses 'good' candidates or pre-optimized configurations. Moreover, although the relevant problem should be three-dimensional, we first investigate the two-dimensional one since it is simpler, and not solved yet.

On a more methodological level, another purpose of the present work is to develop tailored optimization strategies that prove to be efficient in dimension two, and therefore are good candidates to attack the same problem in the physically relevant 3 dimensional case [3, 7]. In particular, this belief motivates the efforts we make on the design of dedicated mutation and cross-over operators for the genetic algorithm.

The article is organized as follows. In Section 2, we first introduce the models : the first one (a Thomas-Fermi type model) is of quantum nature, but can indeed be recasted in the form of a classical one, the second one (ruled by a two body Lennard-Jones interaction potential) genuinely is of classical nature. The crucial difference

between the two is that the former does not allow for binding of particles, whereas the latter does. Therefore, we need to make use of compactification strategies for the former. These strategies make the numerics simpler in some sense, and we then see that deterministic optimization algorithms allow us to find satisfactory result, at the price of some postprocessing extra-work. On the contrary, in the case of the Lennard-Jones potential, the minimization problem can be directly attacked, without compactification, but then requires global optimization strategies, that we here choose to be genetic like algorithms. Let us point out that a wide literature exists on the minimization of Lennard-Jones clusters, but as emphasized above, our approach is different. In Section 3, we therefore successively explain the deterministic and stochastic numerical methods we have used, together with the results they provide.

8.2 Presentation of the models

We begin this section by introducing the Thomas-Fermi (TF) model, showing that it can in fact be recast in a two-body model in the special case of dimension two. This is a standard remark, already made in [5]. Next, we introduce different two-body models.

8.2.1 The Thomas-Fermi model in dimension 2

The Thomas-Fermi [15] model treats the nuclei classically, while the electrons are considered as quantum particles defined by their density $\rho \geq 0$. To a system consisting of N identical nuclei at positions X_j (the nuclei are considered as point particles for simplicity ; they are supposed to be of the same charge, here normalized to one), and N electrons with total density ρ, one associates the TF energy :

$$\begin{aligned} E^{\mathrm{TF}}(\rho, \{X_i\}) &= \int_{\mathbf{R}^d} \rho^p + \frac{1}{2} \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \rho(x) V(x-y) \rho(y) dx dy \\ &- \sum_{j=1}^{N} \int_{\mathbf{R}^d} \rho(x) V(x - X_j) dx + \frac{1}{2} \sum_{j \neq k} V(X_j - X_k). \end{aligned} \tag{8.1}$$

The interaction potential V is the Coulomb $d-$dimensional potential ($V(x) = \frac{1}{|x|}$ if $d = 3$, and $V(x) = -\log(|x|)$ if $d = 2$.) The power p in the first term is equal to $\frac{d+2}{d}$ [21]. This term is an approximation of the kinetic energy of the electronic cloud. This approximation is validated through a high density limit [15].

We recall that, in the case we are interested in, the system is neutral, so that the number of electrons is exactly equal to N :

$$\int_{\mathbf{R}^d} \rho = N.$$

As announced, we work here in dimension 2 and are thus dealing with

$$I_N^{\mathrm{TF}} = \inf\Big\{ E^{\mathrm{TF}}(\{X_j\}), \quad X_j \in \mathbf{R}^2 \Big\}, \tag{8.2}$$

where

$$E^{\mathrm{TF}}(\{X_j\}) = \inf\Big\{ E^{\mathrm{TF}}(\rho, \{X_i\}), \quad \rho \geq 0, \quad \rho \in L^1(\mathbf{R}^2) \cap L^2(\mathbf{R}^2), \quad \int_{\mathbf{R}^2} \rho = N \Big\}, \tag{8.3}$$

and

$$\begin{aligned} E^{\mathrm{TF}}(\rho, \{X_i\}) &= \int_{\mathbf{R}^2} \rho^2 - \frac{1}{2} \int_{\mathbf{R}^2} \int_{\mathbf{R}^2} \rho(x) \log|x-y| \rho(y) dx dy \\ &+ \sum_{j=1}^{N} \int_{\mathbf{R}^2} \rho(x) \log|x - X_j| dx - \frac{1}{2} \sum_{j \neq k} \log|X_j - X_k|. \end{aligned} \tag{8.4}$$

Note that the energy is quadratic with respect to ρ, so that the Euler-Lagrange equation of the minimization problem (8.3) is linear. More precisely, this equation reads, at the minimizer $\bar{\rho}$:

$$2\bar{\rho} - \bar{\rho} * \log|x| + \sum_{j=1}^{N} \log|x - X_j| = \theta, \tag{8.5}$$

where θ is the Lagrange multiplier associated to the constraint $\displaystyle\int_{\mathbf{R}^2} \rho = N$. Taking the Laplacian of this equation, we have :

$$-\Delta\bar{\rho} + \pi\bar{\rho} = \pi \sum_{j=1}^{N} \delta_{X_j}. \tag{8.6}$$

We therefore introduce the Yukawa potential W_{TF} of parameter $\sqrt{\pi}$, i.e the solution going to zero at infinity of

$$-\Delta W_{\mathrm{TF}} + \pi W_{\mathrm{TF}} = \delta_0. \tag{8.7}$$

Let us point out that the potential W_{TF} is in fact equal to

$$W_{\mathrm{TF}}(x) = \frac{1}{2} K_0(\sqrt{\pi}|x|), \tag{8.8}$$

where K_0 is the modified Bessel function, as defined in [1]. In particular, it is a radial decreasing function.

One now easily deduces that the minimizing electronic density $\bar{\rho}$ reads

$$\bar{\rho}(x) = \pi \sum_{j=1}^{N} W_{\mathrm{TF}}(x - X_j). \tag{8.9}$$

Next, going back to the expression of the energy, and using (8.5) and (8.9), we have :

$$\begin{aligned} E^{\mathrm{TF}}(\{X_i\}, \bar{\rho}) &= \int_{\mathbf{R}^2} \bar{\rho}^2 + \frac{1}{2} \int_{\mathbf{R}^2} \bar{\rho}(x) \Big(\sum_{i=1}^{N} \log|x - X_i| - \bar{\rho} * \log|x| \Big) dx \\ &\quad + \frac{1}{2} \sum_{j=1}^{N} \Big(\int_{\mathbf{R}^2} \bar{\rho}(x) \log|x - X_j| dx - \sum_{i \neq j} \log|X_i - X_j| \Big) \\ &= \int_{\mathbf{R}^2} \bar{\rho}^2 + \frac{1}{2} \int_{\mathbf{R}^2} \bar{\rho}(-2\bar{\rho} + \theta) + \frac{1}{2} \sum_{j=1}^{N} \big(2\bar{\rho} - \theta + \log(|\cdot - X_j|)\big)(X_j) \\ &= \sum_{j=1}^{N} \big(\bar{\rho} + \frac{1}{2} \log(|\cdot - X_j|)\big)(X_j) \\ &= \sum_{i \neq j} W_{\mathrm{TF}}(X_i - X_j) + N \lim_{x \to 0} \big(W_{\mathrm{TF}}(x) + \frac{1}{2} \log(|x|) \big). \end{aligned} \tag{8.10}$$

Note that the limit appearing above does exist due to the definition of W_{TF} and is of course a constant, independent of N and of X_j, denoted by A. The main consequence is that the TF energy is in fact a two-body energy :

$$E^{\mathrm{TF}}(\{X_i\}) = \sum_{j \neq i} W_{\mathrm{TF}}(X_i - X_j) + NA.$$

The constant A being independent of N and on X_j. This is why we focus on two-body models in this article. As A does not affect the minimization with respect to X_i, we make the slight abuse of forgetting it from now on.

Since W_{TF} is radially symmetric and decreasing, it immediately follows that the minimization problem (8.2) has no solution, the minimum being reached only when all X_i go to infinity. This no-binding property of the Thomas-Fermi model is well known [15]. Consequently, we now slightly modify problem (8.2) in order to have it well-defined, following [19, 20, 27]. Given a periodic lattice ℓ, we consider the problem of finding the minimum energy configuration subject to ℓ-periodic boundary conditions. More precisely, we set :

$$I_N^{\mathrm{TF}}(\ell) = \inf \Big\{ E^{\mathrm{TF}}_{\sqrt{N}\ell}(\{X_j\}), \quad X_j \in Q(\sqrt{N}\ell), \Big\}, \tag{8.11}$$

where $Q(\sqrt{N}\ell)$ is the primitive unit cell of the lattice $\sqrt{N}\ell$ (and which could be replaced by any unit cell of the lattice), and the energy $E^{\rm TF}_{\sqrt{N}\ell}$ is defined as follows :

$$E^{\rm TF}_{\sqrt{N}\ell}(\{X_i\}) = \frac{1}{2}\sum_{j\neq i}\sum_{k\in\sqrt{N}\ell} W_{\rm TF}(X_i - X_j + k) = \frac{1}{2}\sum_{j\neq i}\sum_{k\in\ell} W_{\rm TF}(X_i - X_j + \sqrt{N}k). \tag{8.12}$$

This energy is exactly equal to the average energy of an infinite set of atoms with positions $\{X_j + k,\, 1 \leq j \leq N,\, k \in \sqrt{N}\ell\}$. The fact that we scale the lattice by a factor $\sqrt{N}$ corresponds to fixing the minimum *atomic density*, i.e the minimum average number η of atoms per unit volume.

This periodization allows us to deal with a well-posed problem (8.11), since the domain in which the positions X_j vary is compact, ensuring the existence of a solution to (8.11).

The choice of the lattice ℓ is rather arbitrary. We choose in the sequel the hexagonal lattice, since it is the minimizing lattice of fixed atomic density (see subsection 8.3.1 below).

8.2.2 Two-body models

As we have just seen, considering the 2-dimensional Thomas-Fermi model amounts to considering in fact a 2-body interaction potential between the nuclei, and therefore to minimizing

$$I^{\rm TF}_N(\ell) = \inf\left\{\frac{1}{2}\sum_{j\neq i}\sum_{k\in\sqrt{N}\ell} W_{\rm TF}(X_i - X_j + k), \quad X_i \in \sqrt{N}Q(\ell)\right\}, \tag{8.13}$$

where ℓ is the hexagonal lattice, $Q(\ell)$ one of its unit cell, and $W_{\rm TF}$ is defined by (8.8). We shall investigate in the next section the behaviour of the minimizer to (8.13) as N grows to infinity.

Alternatively, we wish to consider another type of 2-body interaction potential, that, because of its genuine confining properties, does not require the periodization trick above to give rise to a well-posed minimization problem. As a toy-model for such a potential, we consider the famous Lennard-Jones potential

$$W_{\rm LJ}(x) = \frac{1}{|x|^{12}} - \frac{2}{|x|^6}. \tag{8.14}$$

This potential is well-known, for instance, to satisfyingly model noble gases [14]. It is a radially symmetric function, decreasing with respect to $|x|$ if $0 < |x| < 1$, and increasing if $|x| > 1$.

For such a potential, we may therefore legitimately consider the minimizer of

$$I^{LJ}_N = \inf\left\{\frac{1}{2}\sum_{j\neq i} W_{\rm LJ}(X_i - X_j), \quad X_i \in \mathbf{R}^2\right\}, \tag{8.15}$$

and investigate (numerically) its behaviour as N grows.

8.3 Numerical strategies and results

8.3.1 Minimizing over periodic lattices

As a preliminary, we first indicate results on the problem of minimizing the TF energy of a periodic lattice under the constraint of fixed atomic density (or equivalently atomic density bounded from above). Such a work is necessary in order to determine which of the periodic lattices is the good candidate to be the global minimizer of the energy when the assumptions of periodicity is relaxed.

Let W_{TF} be the interaction potential defined by (8.8). Let us denote by

$$E^{\mathrm{TF}}(\ell) = \sum_{k\in\ell\setminus\{0\}} W_{\mathrm{TF}}(k), \tag{8.16}$$

the energy of the lattice ℓ. It is easily seen that $E^{\mathrm{TF}}(\ell)$ is also the average energy of the lattice ℓ in the following sense :

$$E^{\mathrm{TF}}(\ell) = \lim_{R\to\infty}\left(\frac{1}{\#(\ell\cap B_R)}\sum_{p\in\ell\cap B_R}\sum_{p\neq q\in\ell\cap B_R} W_{\mathrm{TF}}(p-q)\right).$$

The problem under consideration here is the following : fix an atomic density $\eta > 0$, or equivalently a volume $V = \frac{1}{\eta}$, and minimize the energy (8.16) over the set of all periodic lattice ℓ satisfying, if $Q(\ell)$ is its unit cell, $|Q(\ell)| = V = \frac{1}{\eta}$. In other words, find a solution to

$$I^{\mathrm{TF}}_{\mathrm{per}}(\eta) = \inf\left\{E^{\mathrm{TF}}(\ell), \quad |Q(\ell)| = V = \frac{1}{\eta}\right\}. \tag{8.17}$$

Note that, since the interaction potential W_{TF} (8.8) is radially symmetric and decreasing, the minimization problem (8.17) is equivalent to minimizing the energy under the constraint $|Q(\ell)| \leq V$. In other words, fixing the atomic density amounts to fixing a minimum atomic density.

The first task in numerically computing this minimum is the parameterization of the set on which we minimize, that is, the set of lattices of atomic density η. A lattice may be defined by any of its basis, that is, vectors (a, b) such that

$$\ell = \big\{ia + jb, \quad i \in \mathbf{Z}, \quad j \in \mathbf{Z}\big\}.$$

We recall the simple

Theorem 8.3.1 (Engel, [8]) *For any periodic lattice $\ell \subset \mathbf{R}^2$, there exists a basis (a, b) of ℓ such that :*

$$\begin{cases} |a| \leq |b|, \\ \widehat{(a,b)} \in [\frac{\pi}{3}, \frac{\pi}{2}], \end{cases} \tag{8.18}$$

where $\widehat{(a,b)}$ denotes the angle between a and b.

Since the volume $|a \wedge b|$ of the unit cell is supposed to be equal to V, the lattice ℓ is entirely defined by the angle $\theta = \widehat{(a,b)}$ and the length $x = |a|$. In addition, the inequality $|a| \leq |b|$ implies $0 \leq x \leq \frac{1}{\sqrt{\sin\theta}}$. Therefore, let us define

$$\mathcal{A} = \left\{ (x,\theta), \quad \frac{\pi}{3} \leq \theta \leq \frac{\pi}{2}, \quad 0 \leq x \leq \frac{1}{\sqrt{\eta \sin\theta}} \right\}.$$

The set $\mathcal{A}$ is in bijection with the set $\mathcal{L}_\eta$ of periodic lattice of atomic density η through the application

$$\begin{array}{rccl} \Phi : & \mathcal{A} & \longrightarrow & \mathcal{L}_\eta \\ & (x,\theta) & \longmapsto & \left\{ \begin{pmatrix} ix + \frac{j}{\eta x \tan\theta} \\ \frac{j}{\eta x} \end{pmatrix}, \quad \begin{pmatrix} i \\ j \end{pmatrix} \in \mathbf{Z}^2 \right\}. \end{array}$$

Hence, the minimization problem (8.17) may be reduced to a minimization on the subset $\mathcal{A}$ of $\mathbf{R}^2$, which is compact (this property ensures the existence of a minimum). In this parameterization, the hexagonal lattice of atomic density η is defined by the values $(x,\theta) = \left(\sqrt{\frac{2}{\eta\sqrt{3}}}, \frac{\pi}{3}\right)$.

Table 8.1 gives the results of calculations performed on this problem, with $\eta = 1$. The initial guess is chosen randomly in $\mathcal{A}$. The calculation has been performed both with the built-in optimization toolbox of Matlab [18] and an in-house developped Fortran code. In either case, the algorithm is a first order algorithm (that performs very well in this case when the minimization space is of a low dimension), namely a Polak-Ribière non linear conjugated gradient algorithm with Wolfe or Goldstein-Price line-search (for a complete presentation of these algorithms see [6, Part 1]). Both options give similar results, displayed in table 8.1.

In table 8.1, the error is evaluated from the quantity $\max(|x - \sqrt{\frac{2}{\eta\sqrt{3}}}|, |\theta - \frac{\pi}{3}|)$, which is the distance between the computed minimizer (x,θ) and the hexagonal lattice $\left(\sqrt{\frac{2}{\eta\sqrt{3}}}, \frac{\pi}{3}\right)$ in $\mathcal{A}$, which is the natural candidate for the optimal configuration. We observe that, independently from the initial guess, the minimization procedure indeed converges to this particular hexagonal lattice.

8.3.2 The Thomas-Fermi case with periodic boundary conditions

Keeping in mind the results of the previous section, which show that for a fixed atomic density, the lattice with minimum energy is the hexagonal lattice, we now turn to the Thomas-Fermi case with periodic boundary conditions.

In view of the previous section, we fix the lattice H to be the hexagonal lattice with unit length, that is :

$$H = \left\{ \begin{pmatrix} i \\ \frac{1}{2}i + \frac{\sqrt{3}}{2}j \end{pmatrix}, \quad i,j \in \mathbf{Z} \right\}.$$

Initial Guess $\begin{pmatrix} x \\ \theta \end{pmatrix}$	$\begin{pmatrix} 0.1492 \\ 1.1534 \end{pmatrix}$	$\begin{pmatrix} 0.2135 \\ 1.3633 \end{pmatrix}$	$\begin{pmatrix} 0.2925 \\ 1.1513 \end{pmatrix}$	$\begin{pmatrix} 0.0164 \\ 1.4382 \end{pmatrix}$
Error	1.7×10^{-5}	1.6×10^{-5}	2.0×10^{-5}	1.8×10^{-5}
Gradient norm	1.5×10^{-10}	9.7×10^{-11}	1.5×10^{-10}	1.2×10^{-10}
CG iterations	22	22	20	28
Initial Guess $\begin{pmatrix} x \\ \theta \end{pmatrix}$	$\begin{pmatrix} 0.4783 \\ 1.5351 \end{pmatrix}$	$\begin{pmatrix} 0.5007 \\ 1.2664 \end{pmatrix}$	$\begin{pmatrix} 0.9093 \\ 1.3222 \end{pmatrix}$	$\begin{pmatrix} 0.2178 \\ 1.3991 \end{pmatrix}$
Error	2.1×10^{-5}	1.2×10^{-5}	2.1×10^{-5}	1.8×10^{-5}
Gradient norm	1.7×10^{-10}	5.6×10^{-11}	1.7×10^{-10}	1.3×10^{-10}
CG iterations	19	19	14	21

TAB. 8.1 – Numerical results of problem (8.17) obtained by conjugate gradient algorithm (from the optimization toolbox of Matlab). The initial guess is randomly chosen in $\mathcal{A}$.

We define a particular unit cell $Q(H)$ of H :

$$Q(H) = \left\{ \begin{pmatrix} x \\ \frac{1}{2}x + \frac{\sqrt{3}}{2}y \end{pmatrix}, \quad x, y \in [0,1) \right\}.$$

Let N be an integer, which will be the number of atoms per cell. We assume that

$$N = P^2$$

for some integer P. Defining a set of positions $\{X_i\}_{1 \le i \le N}$ such that

$$\forall i \in \{1, 2, \ldots, N\}, \quad X_i \in PQ(H),$$

the corresponding TF energy with PH boundary conditions is

$$E_{PH}^{\mathrm{TF}}(\{X_i\}_{1 \le i \le N}) = \frac{1}{2} \sum_{i \neq j} \sum_{k \in H} W_{\mathrm{TF}}(X_i - X_j + Pk),$$

as defined in (8.12). And the minimization problem we are dealing with here is defined by (8.11) :

$$I_N^{\mathrm{TF}}(H) = \inf \left\{ E_{PH}^{\mathrm{TF}}(\{X_i\}_{1 \le i \le N}), \quad X_j \in PQ(H) \right\}. \tag{8.19}$$

We have performed the calculation of the solution of this minimization problem using the Quasi-Newton (BFGS) algorithm of the built-in Matlab optimization toolbox, adding the constraint that the positions X_i should stay in the unit cell PQ of the lattice PH (this of course is not a limitation, and only enhances the stability of the

Number of atoms	4	9	16	36
Error	2.1×10^{-5}	8.3×10^{-6}	2.5×10^{-5}	4.7×10^{-5}
Gradient norm	9.7×10^{-6}	5.1×10^{-6}	7.1×10^{-6}	8.4×10^{-6}
BFGS iterations	27	36	55	109
Number of atoms	49	64	81	100
Error	1.9×10^{-5}	5.0×10^{-5}	2.2×10^{-6}	8.4×10^{-6}
Gradient norm	8.1×10^{-6}	9.1×10^{-6}	6.9×10^{-7}	5.4×10^{-6}
BFGS iterations	94	143	162	162

TAB. 8.2 – Numerical results of problem (8.19) obtained by the BFGS algorithm (from the optimization toolbox of Matlab).

calculation), and starting from a randomly chosen initial guess. Table 8.2 shows the results of these calculations for $N = 2^2$ up to $N = 10^2$. The error is defined as :

$$\text{Error} = \max\{d(X_i, H),\ 1 \leq i \leq N\},$$

where $d(x, H) = \inf\{\|x - k\|,\ k \in H\}$ denotes the distance between x and the set H.

The algorithm converges in a quite satisfactory way, showing that, as expected, the periodic arrangement is the limit.

8.3.3 The Lennard-Jones potential : unconstrained minimization

Before attacking the case of the Lennard-Jones potential, we need two theoretical results. This case actually is slightly more demanding as there is no enforced confinement of the atoms, for this confinement is indeed built-in in the large scale behaviour of the interaction potential. Therefore, we must carefully choose the initial guess, and continuously control, during the minimization procedure, the positions of the atoms. In order to do that in a non-biased way, we need to understand more quantitatively how the distance between atoms behaves.

8.3.3.1 Theoretical results

We recall that the problem we are dealing with is (8.15), with the interaction potential defined by (8.14).

The first point is, the atoms of a minimizing configuration are isolated from each other :

Theorem 8.3.2 *Let N be a positive integer, and let $\{X_i\}_{1 \leq i \leq N}$ be a solution of (8.15). Then there exists $a > 0$ such that*

$$\forall i \neq j, \quad \|X_i - X_j\| \geq a. \tag{8.20}$$

Moreover, $a \geq 0.7286$.

Let us mention that such bounds from below have already been obtained in the literature [30] : a fixed bound $a = 0.5$, or bounds depending on N (and going to zero as N goes to infinity).

We now give a few definitions : for all $i \in \{1, 2, \dots N\}$, we define the *individual energy* of particle i as

$$E_i(\{X_j\}) = \sum_{j \neq i} W_{\rm LJ}(X_i - X_j). \tag{8.21}$$

Note that the total energy satisfies the following equality :

$$E^{W_{\rm LJ}}(\{X_i\}) = \sum_{1 \leq i < j \leq N} W_{\rm LJ}(X_i - X_j) = \frac{1}{2} \sum_{i=1}^{N} E_i(\{X_j\}).$$

We also denote by $r_{\rm min}(i)$ the distance between X_i and its nearest neighbor, and by $r_{\rm min}$ the overall minimum distance :

$$r_{\rm min}(i) = \inf\{\|X_i - X_j\|, \quad j \neq i\}, \qquad r_{\rm min} = \inf\{r_{\rm min}(i), \quad 1 \leq i \leq N\}.$$

Proof : We will divide the proof of this theorem into several steps :

Step one : For any minimizing configuration $\{X_i\}$, we have

$$\forall i \in \{1, 2, \dots N\}, \quad E_i(\{X_j\}) < -1.$$

Indeed, assume that for some i, we have $E_i \geq -1$. Reordering the particles if necessary, we may assume that X_1 is the particle with lowest first coordinate among $\{X_j, j \neq i\}$. Therefore, if we define a new configuration $\{Y_j\}_{1 \leq j \leq N}$ by $Y_j = X_j$ if $j \neq i$ and $Y_i = X_1 - e_1$, where e_1 is the first vector of the canonical basis, we have, for this new configuration,

$$E_i(\{Y_j\}) = W_{\rm LJ}(1) + \sum_{j \neq 1, i} W_{\rm LJ}(X_j - Y_i) < W_{\rm LJ}(1) = -1.$$

Therefore, computing the energy difference, we have

$$E^{W_{\rm LJ}}(\{Y_j\}) - E^{W_{\rm LJ}}(\{X_j\}) = E_i(\{Y_j\}) - E_i(\{X_j\}) < 0.$$

This is in contradiction with the fact that $E^{W_{\rm LJ}}(\{X_j\})$ is a minimum.

Step two : The minimum distance between two particles is bounded below by $a = 0.518$.

Let i_0 be an index satisfying $r_{\rm min} = r_{\rm min}(i_0)$. For any $k \in \mathbf{N}$, we define

$$\mathcal{N}_k = \{j \neq i_0 \,/\, k r_{\rm min} \leq \|X_{i_0} - X_j\| < (k+1) r_{\rm min}\}, \quad \text{and} \quad N_k = \#\mathcal{N}_k.$$

Since we know that for all $i \neq j$, $\|X_i - X_j\| \geq r_{\min}$, the balls $B_{\frac{r_{\min}}{2}}(X_j)$ do not intersect. We also have, by definition of $\mathcal{N}_k$,

$$\bigcup_{j\in\mathcal{N}_k} B_{\frac{r_{\min}}{2}}(X_j) \subset B_{(k+\frac{3}{2})r_{\min}}(X_{i_0}) \setminus B_{(k-\frac{1}{2})r_{\min}}(X_{i_0}).$$

Hence, $\sum_{j\in\mathcal{N}_k} |B_{\frac{r_{\min}}{2}}(X_j)| \leq \pi r_{\min}^2\big((k+\frac{3}{2})^2 - (k-\frac{1}{2})^2\big)$, which implies that

$$N_k \leq 16k + 8.$$

We now use this estimate to bound from below the energy $E_{i_0}(\{X_{i_0}\})$:

$$\begin{aligned} E_{i_0}(\{X_{i_0}\}) &= \frac{1}{r_{\min}^{12}} - \frac{2}{r_{\min}^{6}} + \sum_{k\geq 1}\sum_{j\in\mathcal{N}_k} W_{\mathrm{LJ}}(X_{i_0} - X_j) \\ &\geq \frac{1}{r_{\min}^{12}} - \frac{2}{r_{\min}^{6}} - \sum_{k\geq 1} \frac{2N_k}{k^6 r_{\min}^6} \\ &\geq \frac{1}{r_{\min}^{12}} - \frac{2}{r_{\min}^{6}} - \left(\sum_{k\geq 1} \frac{16(2k+1)}{k^6}\right)\frac{1}{r_{\min}^6}. \end{aligned}$$

We denote by α the constant $\sum \frac{16(2k+1)}{k^6}$, and set $t = \frac{1}{r_{\min}^6}$. Then, using the fact that $E_{i_0}(\{X_{i_0}\}) < -1$, we have :

$$1 - (2+\alpha)t + t^2 \leq 0.$$

This implies that t is between the zeros of the polynomial $X^2 - (2+\alpha)X + 1$, and in particular that $t \leq \frac{2+\alpha+\sqrt{(2+\alpha)^2-4}}{2}$. Thus,

$$r_{\min} \geq \left(\frac{2}{2+\alpha+\sqrt{(2+\alpha)^2-4}}\right)^{1/6}.$$

Numerical computation of this value gives $r_{\min} \geq 0.5185415283$.

Step three : The minimum distance between two particles is bounded below by $a = 0.7286$

We repeat here the same kind of argument as above : here again, i_0 denotes an index satisfying $r_{\min}(i_0) = r_{\min}$. The integers N_k are defined in the same way as above, and noticing that the interaction potential is an increasing function of the distance r as far as $r \geq 1$, we deduce, using the inequality $2r_{\min} \geq 1$,

$$E_i(\{X_i\}) \geq W_{\mathrm{LJ}}(r_{\min}) - (N_1 - 1) + \sum_{k\geq 2} N_k W_{\mathrm{LJ}}(k r_{\min}).$$

Using the fact that $N_k \leq 16k+8$, we thus have :

$$22 \geq \frac{1}{r_{\min}^{12}} - \frac{2}{r_{\min}^{6}} + \frac{1}{r_{\min}^{12}} \sum_{k\geq 2} \frac{16k+8}{k^{12}} - \frac{1}{r_{\min}^{6}} \sum_{k\geq 2} \frac{32k+16}{k^{12}}.$$

Hence, setting $P = \sum_{k\geq 2} \frac{16k+8}{k^{12}}$ and $Q = \sum_{k\geq 2} \frac{32k+16}{k^{12}}$, and $t = \frac{1}{r_{\min}^{6}}$, we have $(P+1)t^2 - (Q+2)t - 22 \leq 0$, which implies $t \leq \frac{Q+2+\sqrt{(Q+2)^2+88(P+1)}}{2P+2}$, so that

$$r_{\min} \geq \left(\frac{2P+2}{Q+2+\sqrt{(Q+2)^2+88(P+1)}} \right)^{1/6}.$$

Numerical evaluation of this quantity gives $r_{\min} \geq 0.7286006078$, which concludes the proof of Theorem 8.3.2. □

We now turn to the problem of establishing an upper bound on the distance between two particles. Although it is commonly admitted that this bound should be of order $N^{1/2}$ (and $N^{1/3}$ in dimension 3), this fact remains to be rigorously proved, to the best of our knowledge. We provide here a very crude bound in this respect :

Theorem 8.3.3 *Let N be a positive integer, and let $\{X_i\}_{1\leq i\leq N}$ be a solution of (8.15). Then, we have :*

$$\forall i,j \in \{1,2,\ldots,N\}, \quad \|X_i - X_j\| \leq N. \tag{8.22}$$

Proof : Arguing by contradiction, we assume that $\max \|X_i - X_j\| > N$. Without loss of generality, we may assume that this maximum is reached for $i=1$ and $j=N$. In addition, since the energy is invariant under rotations and translations, we may assume that $X_1 = 0$ and that the abscissa of X_N is zero. We define, for any $k \in \mathbf{N}$,

$$L_k = \{x \in \mathbf{R}^2,\ k \leq x \cdot e_1 < k+1\}.$$

Let P be the smallest integer such that $\|X_N\| < P+1$. We know that $P \geq N$, and that

$$\forall i \in \{1,2,\ldots,N\}, \quad X_i \in \bigcup_{0\leq k\leq P} L_k.$$

Hence, there exists at least an integer k satisfying $1 \leq k \leq P-1$ such that $L_k \cap \{X_i\} = \emptyset$. Fixing such a k, we may reorder the X_i's so that there exists $N_1 < N$ such that

$$\forall i \leq N_1,\ X_i \in \bigcup_{p<k} L_p \quad \text{and} \quad \forall i > N_1,\ X_i \in \bigcup_{p>k} L_p.$$

We now define a new configuration $\{Y_i\}$ as follows (e_1 is the first vector of the canonical basis of $\mathbf{R}^2$) :

- if $i \leq N_1$, then $Y_i = X_i$;
- if $i > N_1$, then $Y_i = X_i - \alpha e_1$ where $\alpha = |(X_{N_1+1} - X_{N_1}) \cdot e_1| - 1 > 0$.

We now compute the energy of this configuration $\{Y_i\}$:

$$\begin{aligned} E(\{Y_i\}) &= \sum_{1\le i<j\le N} W_{\mathrm{LJ}}(Y_i - Y_j) \\ &= \sum_{1\le i<j\le N_1} W_{\mathrm{LJ}}(X_i - X_j) + \sum_{N_1+1\le i<j\le N} W_{\mathrm{LJ}}(X_i - X_j) \\ &\quad + \sum_{1\le i\le N_1<j\le N} W_{\mathrm{LJ}}(X_i - X_j - \frac{1}{2}e_1) \\ &< \sum_{1\le i<j\le N_1} W_{\mathrm{LJ}}(X_i - X_j) + \sum_{N_1+1\le i<j\le N} W_{\mathrm{LJ}}(X_i - X_j) \\ &\quad + \sum_{1\le i\le N_1<j\le N} W_{\mathrm{LJ}}(X_i - X_j) = E(\{X_i\}), \end{aligned}$$

since the function $W_{\mathrm{LJ}}(x)$ increases with $\|x\|$ when $\|x\| \ge 1$. We thus reach a contradiction, proving (8.22). □

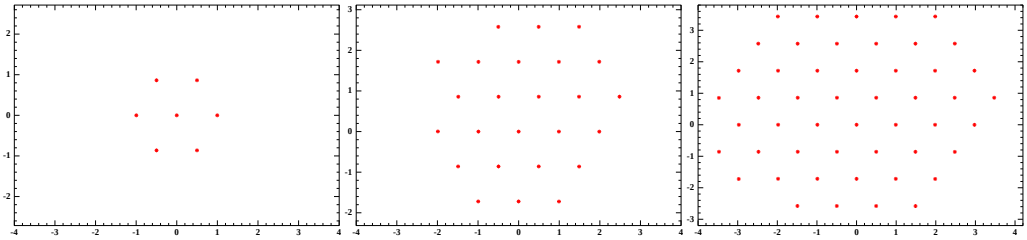
With the help of the above two results, we may now attack the direct numerical minimization. Theorem 8.3.2 ensures that for a configuration with minimal energy, the particles are uniformly separated from one another. Hence, their interaction energy is bounded from above. In addition, the initial guess for the minimization may be chosen accordingly. On the other hand, Theorem 8.3.3 helps in determining the size of the box where all atoms are to be kept.

8.3.3.2 Construction of a reference configuration

In section 8.3.1 we have seen that some hexagonal lattice is a good candidate to be the global minimizer for an infinite number set of atoms. As for the numerical experiments we deal with a finite number of atoms, we will construct for a given number N of atoms a reference configuration based on such an hexagonal lattice. The reference configuration, denoted by $\mathcal{C}^{ref}$, is obtained by truncating an infinite hexagonal lattice with unit length to an hexagonal configuration with only N atoms and then relaxing it using the conjugate gradient algorithm. More precisely, we use the following procedure :

1. Place the first atom X_1 on a node of the hexagonal lattice.
2. For i from 2 to N do
   2.1 Find the best position for X_i over the hexagonal lattice which minimizes the energy of the configuration $\{X_j\}_{1\le j\le i}$.
3. Perform a conjugate gradient on the configuration $\{X_j\}_{1\le j\le N}$: this gives the reference configuration $\mathcal{C}^{ref}$.

Note that the configuration obtained is still periodic which means, in particular, that this configuration is at least a local minimum. In the following we will compare the results we will obtain to this reference configuration. Figure 8.1 shows some reference configurations obtained by the procedure below.



$N = 7$ $N = 25$ $N = 50$

FIG. 8.1 – Different reference configurations for $N = 7$, 25 and 50.

8.3.3.3 Deterministic techniques

We present in this section some numerical results on the minimization problem (8.15), using the same techniques as those of section 8.3.2. The initial guess is randomly chosen under the constraint $a \leq |X_i - X_j| \leq N$ for $1 \leq i < j \leq N$ and where a is given by theorem 8.3.2. We shall see that, here, these techniques do not provide sufficiently satisfactory results. Basically, the deterministic algorithms converge to configurations that are locally not periodic, but are very close to be a subset of the hexagonal lattice H. Furthermore, these configurations are not global minimizers, for their energy can be further decreased by simple manipulations. This therefore justifies the need to resort to other optimization strategies, that will be examined in the next section.

When a standard optimization algorithm (such as a conjugate gradient or a quasi-Newton method) is performed, starting from a position satisfying (8.20) and (8.22), we observe that the algorithm stops at a configuration where the particles tend to cluster into small groups of a few particles, each cluster being far from each other. The potential being weak at infinity, such a configuration involving clusters of particles is indeed (numerically) a stationary point of the energy. Figure 8.2 presents an example of such a configuration. Other initial guesses, and other deterministic strategies would lead to different configurations, however exhibiting the same qualitative behaviour.

Figure 8.2 can therefore be considered as a prototype for the output of a deterministic algorithm in this setting. This might look very disappointing at first sight, but one point should be made : a close-up would reveal that the small clusters look very much like subsets of an hexagonal lattice. Based on this observation, we decide to change the "global" components of the structure, leaving the "local" ones unchanged. This gives rise to the following strategy, that we henceforth call the *closing-in* algorithm.

Closing-in Algorithm

1. Perform a conjugate gradient on the configuration $\{X_i\}$;
2. Determine the blocks, and identify B_1 the largest one of them ;
3. Translate B_2, the closest block to B_1, towards B_1 until their distance is 1 ;

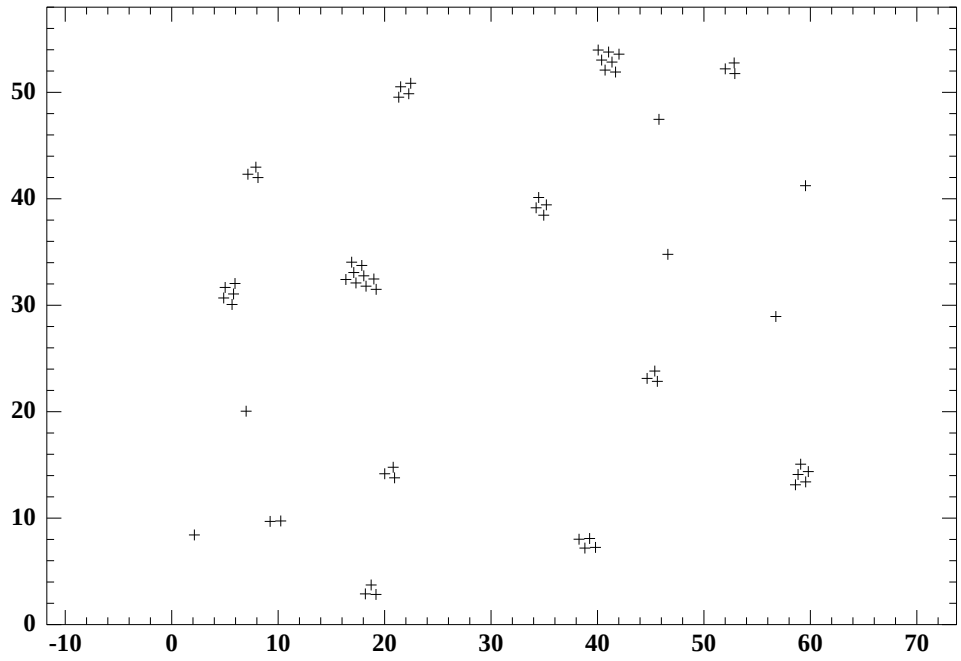


FIG. 8.2 – The result of a conjugate gradient computation for problem (8.13).

4. Compute the energy difference DE between this new configuration and the old one ;
5. If $DE < 0$, adopt the new configuration and go to 1., otherwise keep the old one, and terminate.

Let us detail steps 2 and 3 above :

– in step 2, the algorithm to find a block is the following : a) fix an atom X_i, b) find all the neighbors of X_i, that is, the X_j satisfying $\|X_i - X_j\| \leq r_{\text{neighb}}$, where r_{neighb} is the maximum neighbor distance (r_{neighb} needs to be larger than one, and is taken to be 1.2 in all the examples we give here), and c) find the neighbors of the neighbors, and so on (see figure 8.3).

– In step 3, we first locate the largest block B_1, and then locate the block B_2 which is the closest one to B_1. The distance

$$d(B_1, B_2) = \inf\{\|X_i - X_j\|, \quad X_i \in B_1, \quad X_j \in B_2\},$$

between the blocks satisfy $d(B_1, B_2) > r_{\text{neighb}}$. We fix i and j such that $\|X_i - X_j\| = d(B_1, B_2)$, with $X_i \in B_1$ and $X_j \in B_2$, and translate B_2 of the vector $(1 - \frac{1}{\|X_i - X_j\|})(X_i - X_j)$. This drives X_j to a distance 1 from X_i, and hopefully decreases the energy (in practice it usually does).

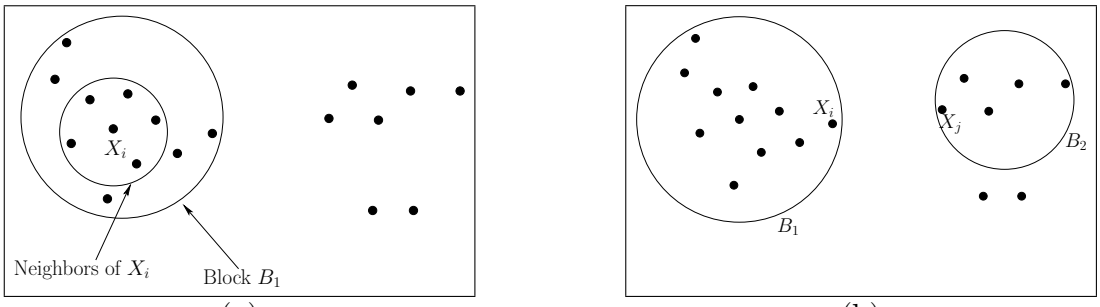


FIG. 8.3 – Illustration of the construction of a "block" (a) and the distance between two blocks (b).

It is to be remarked that the conjugate gradient calculation need not converge within the loop : a stopping criterion "norm of gradient less than 10^{-1}" is usually sufficient. When the iteration terminates, it is useful to perform a more precise conjugate gradient calculation, only for this last iteration.

An example of the result of the "closing-in" algorithm is shown in figure 8.4. Note that there is now only one block of atoms, contrarily to the input configuration of figure 8.2, and they seem to be periodically distributed. Let us argue more quantitatively. We determine among the atoms of our configuration (denoted by $\mathcal{C}$) the ones with the lowest individual energy. Next, we consider two of its neighbors. This defines a periodic cell, of a periodic lattice, denoted by $\mathcal{A}$. We now evaluate the maximum distance between our configuration $\mathcal{C}$ and this periodic lattice $\mathcal{A}$, defined

by

$$\sup_{X_i \in \mathcal{C}} \inf_{Y \in \mathcal{A}} \|X_i - Y\| \tag{8.23}$$

which is about 0.3. In addition, the average distance

$$\frac{1}{N} \sum_{i=1}^{N} \inf_{Y \in \mathcal{A}} \|X_i - Y\| \tag{8.24}$$

is about 0.07, which is small compared to the size of the unit cell of $\mathcal{A}$ which is of the order of 1. However, it seems clear that this configuration is only a local minimum (one may arbitrarily move a nucleus from the boundary and replace it somewhere else in a clever way). We may formulate this observation somewhat vaguely saying that the configuration shown in figure 8.4 is of too a high energy, due to the fact that outer atoms, too numerous in this configuration, have higher individual energies than inner ones.

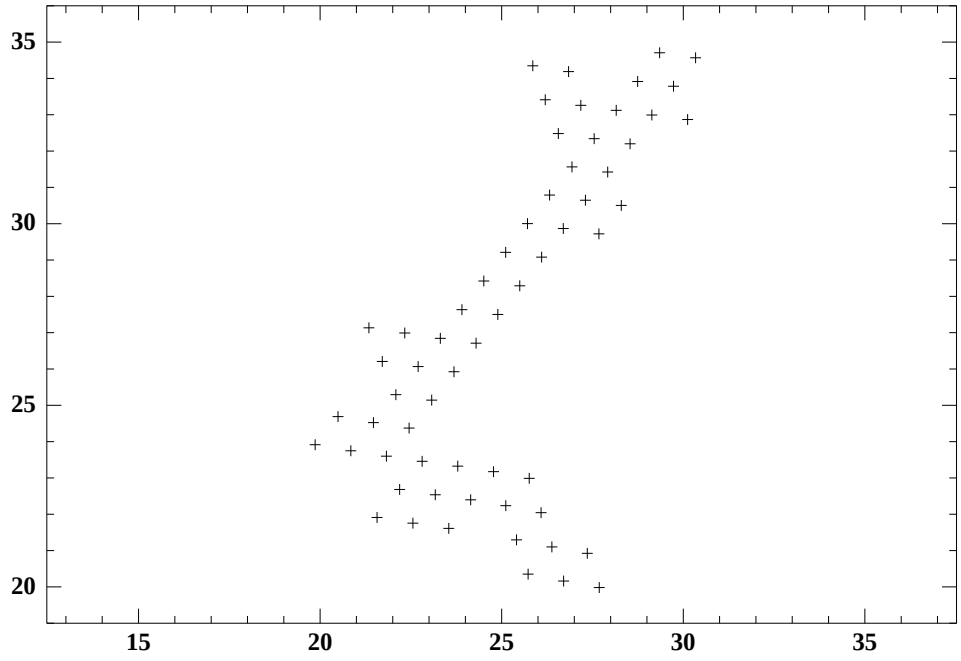


FIG. 8.4 – The result of the "closing-in" algorithm.

A way to further improve the energy is to allow the configuration to be more symmetric. For this purpose, we insert the output of the previous "closing-in" algorithm as an input for the following *symmetrization* algorithm :

Symmetrization algorithm

1. Perform a conjugate gradient calculation on the configuration $\{X_i\}$;

2. For i from 1 to N do
   2.1. Define $\{Y_j\}_{1\leq j\leq 2N}$ to be the configuration of $2N$ particles defined by $Y_j = X_j$ if $j \leq N$, and $Y_j = 2X_i - X_{j-N}$ otherwise ;
   2.2. In this new configuration, compute the individual energy of the atoms ;
   2.3. Delete the N particles with highest energy : this generates a new configuration $\{Z_j\}_{1\leq j\leq N}$;
3. Compute the new energy difference $DE = E(\{Z_i\}) - E(\{X_i\})$;
4. If $DE < 0$, configuration $\{Z_i\}_{1\leq i\leq N}$ replaces the configuration $\{X_i\}_{1\leq i\leq N}$ and goto 1, otherwise keep the old configuration $\{X_i\}_{1\leq i\leq N}$ and terminate.

This algorithm favors symmetric configurations, *provided* the symmetrization diminishes the energy. Applying it to the configuration of figure 8.4, one finds the configuration shown in figure 8.5, that really seems to be periodic. Indeed, computing its distance to the closest periodic lattice (defined by (8.23)), one finds 0.079, which is very small compared to the size of unit cell of the lattice, which is approximately 1. It thus seems that the minimization algorithm consisting of the "closing-in" pro-

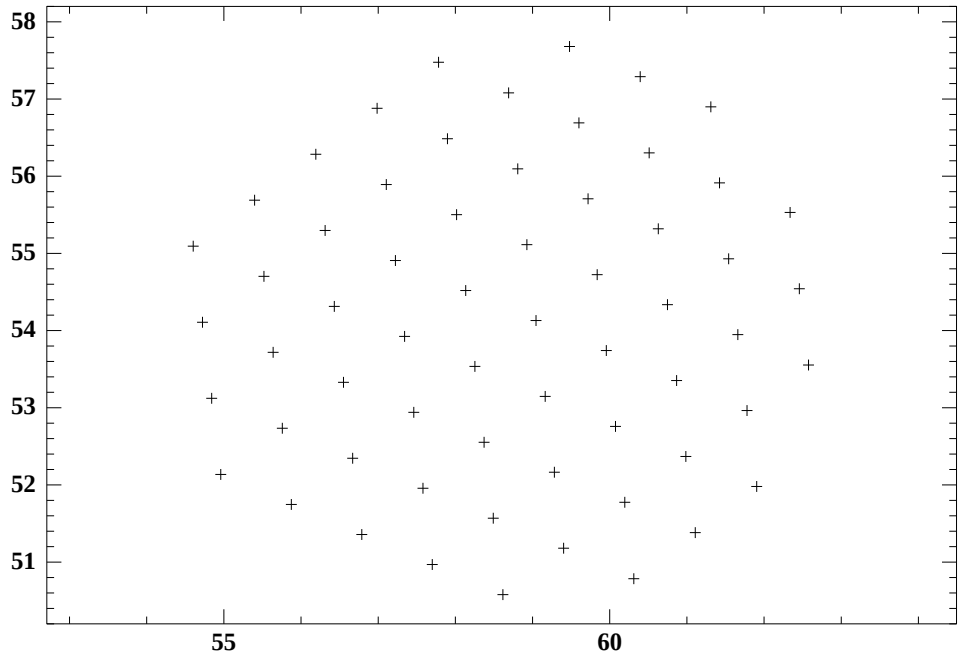


FIG. 8.5 – The result of the symmetrization algorithm.

cedure followed by the "symmetrization" procedure gives a configuration of a very low energy, exhibiting a periodic-like structure. However, one can see on figure 8.5 that it is still possible to decrease the energy by moving some of the atoms to a more appropriate place. For instance, moving the four extreme-left atoms into the empty corners of the hexagon defined by the boundary atoms surely is likely to decrease the energy. This can indeed be checked by computations. Knowing that

this configuration is not a global minimizer, a two-fold question arises : can we have insight on the global minimizer (or at least check that a periodic structure is a good candidate) ? if this (tentative) global minimizer exhibits a periodic structure, is the periodic cell the same as that we have determined in the present section ?

In order to address this question, it is necessary to resort to other types of techniques. We shall use genetic-like algorithms. As will be seen in the next section, the "closing-in" and "symmetrization" procedures that we have developed above as post-processing tools for the deterministic strategies will indeed be still useful in this context, this time in a more systematic way and within the iteration loop (through the specific definition of the parameters of the genetic algorithms).

8.3.3.4 Genetic algorithms

We begin this section by briefly presenting the basic steps of a genetic algorithm (GA). Next, we detail the specific operators we developed for the minimization problem with the Lennard-Jones potential. Note that genetic algorithms have been used in other contexts to solve the Lennard-Jones problem [3, 7].

In order to optimize a given *objective function* f over a given search space $\mathbb{E}$, a genetic algorithm evolves a population of individuals (i.e., a P-uple of points in the search space), usually initialized randomly. The population undergoes some artificial Darwinian evolution based on the *fitness* F of each individual. The fitness of an individual is directly related to the value of the objective function of this individual (a typical example of a fitness function is the objective function itself, but this will not be the case in the implementation we shall make below).
The loop of the algorithm called *a generation* is made up of the following steps :

- *Selection* : the selection operator selects among the parents those who will generate offsprings, the genitors.
- *Creation of new individuals* : by *crossovers* (recombinations of k parents) and *mutations*.
- *Evaluation* : for each offspring the fitness is computed.
- *Replacement* : this operator discriminates among the individuals of the current population those who will be the individuals for the next generation (survival of the fittest).

A basic stopping criterion is when the maximum number of generations fixed by the user is reached.

Genetic algorithms are known to be powerful tools, *provided* they are conveniently adapted to the specific problem under consideration. Otherwise, when utilized as ready -to-use black boxes they often simply give poor results, or even no result at all. The success of such algorithms is indeed intimately linked with a dedicated choice of crossover and mutation operators.

We have therefore made an in-house development [2] of specific operators taking into account the specific properties of our minimization problem, and the ideas presented in the previous sections.

- *The objective function :* an individual $\{X_j\}_{1\leq j\leq N}$ will be replaced by the result of the "closing-in" procedure (of section 8.3.3.3) denoted by $\{Y_j\}_{1\leq j\leq N}$. The

objective function is given by the energy of the new configuration $\{Y_j\}_{1\leq j\leq N}$.

- *The crossover :* two parents $\{X_j^1\}_{1\leq j\leq N}$ and $\{X_j^2\}_{1\leq j\leq N}$ define a configuration of $2N$ particles $\{Y_j\}_{1\leq j\leq 2N}$ where $Y_j = X_j^1$ if $j \leq N$ and $Y_j = X_{j-N}^2$ if $j > N$. After computing the individual energy of the particles in this configuration, we delete the N particles with the highest individual energy. We obtain a configuration $\{Z_j\}_{1\leq j\leq N}$ defining a child which will replace the parent $\{X_j^1\}_{1\leq j\leq N}$.
- *The mutation :* two mutation operators are sequentially used. The first one is based on a symmetrization procedure similar to that presented in the previous section. The symmetry axis is given by two particles X_i and X_j randomly chosen in the configuration $\{X_j\}_{1\leq j\leq N}$. We note that this operator may favor symmetric configurations, and that using other mutation operators (without symmetry) at the beginning of the algorithm lead to the same kind of results. This operator have the advantage to be more efficient and to converge faster. The second mutation operator is used only at the end of the algorithm in order to improve the convergence. Its role is to avoid the local minima with very low energy such as the one given on Figure 8.5. The main idea is to move some particles on the boundary of the configuration near "their appropriate" place. The conjugate gradient will do the rest of the job. This operator proceeds as follows :
  - Delete P particles with the highest individual energy from a configuration $\{X_j\}_{1\leq j\leq N}$.
    The number P is randomly chosen in the range $[1, P_{max}]$ where P_{max} is the number of particles with individual energy higher than -6 (such particles are those defining the boundary of the configuration).
  - Generate Q new particles for some $P \leq Q \leq N + P$.
    A new particle is randomly generated and is only kept if its individual energy is less than $-.5$. More precisely, if the average distance to the closest periodic lattice is small (≤ 0.2 in practice) a new particle is generated by adding zero-mean Gaussian perturbation of standard deviation 0.5 to a node of the closest periodic lattice. If not, the point is randomly generated in a box of size $N \times N$.
  - Perform a conjugate gradient minimization on the newly obtained configuration $\{Y_j\}_{1\leq j\leq N-P+Q}$.
  - Delete the $(Q - P)$ particles with highest individual energy and the configuration obtained now replaces the parent $\{X_j\}_{1\leq j\leq N}$.

Calculations have been performed with this algorithm for the cases of $N = 50, 100, 200$ atoms. The size of the population that is handled is 10. In each case we have improved the results given by the deterministic algorithms in the following sense. In all cases the algorithm ends up with a configuration that has a lower energy than that given by the deterministic algorithms. In the cases $N = 50$ and $N = 100$, the algorithm found the same configuration as the reference configuration $\mathcal{C}^{ref}$ constructed in section 8.3.3.2. In addition, this configuration has the same "periodic structure" in the sense of the construction of formulae (8.23) and (8.24). This confirms the fact that the structure found through a carefully implemented deterministic algorithm is a good candidate to be a global minimizer. Table 8.3 gives the results obtained

N	50	100	200
Energy of $\mathcal{C}^{ref}$	-137.48998	-293.69715	-613.66974
Search space	$[-25, 25]^{100}$	$[-50, 50]^{200}$	$[-100, 100]^{400}$
Symmetrization Alg. : energy found	-136.591574	-292.891543	-612.158901
Symmetrization Alg. : time[1]	$4mn$	$38mn$	$6h\ 23mn$
GA : energy found	-137.48998	-293.69715	-613.05640
GA : time[2]	$7mn$	$25mn$	$> 3\ days$

[1] *CPU time.* [2] *user time on 10 parallel processors.*

Tab. 8.3 – Results obtained by optimization with the symmetrization algorithm and the genetic algorithms.

by the genetic algorithms in these different cases.

We emphasize that most of the CPU time is consumed in the improvement phase : for example, for 100 particles, more than half of the CPU time is used to move from a configuration of energy -292.798189 to the solution with an energy of -293.697155 shown in figure 8.6. This situation is well known when genetic algorithms are applied on problems with many local minima and when some of the minima have an energy close to that of the global minimizer.

8.4 Conclusion

We have presented here numerical results that seem to indicate (in our opinion at least) that, in dimension 2, periodic-like structures are indeed global minimizers of the total energy for various two-body models (with radially symmetric potentials).

The computation time being rather high, we have not been able so far to carry any computation, for a system of reasonable size, in a 3-dimensional setting. However, we believe that the algorithms we have developped here in the 2-dimensional case will be of precious help when tackling the 3D case, which we hope to do in the future.

Let us finally point out that, on the theoretical ground, the crystal problem remains to be solved in dimension higher than or equal to 2, and although the present results seem to indicate that the energy minimum is indeed periodic, it gives no clue about the way to prove it. Even the bounds on the distance between particles shown in section 8.3.3.1 are far from optimal.

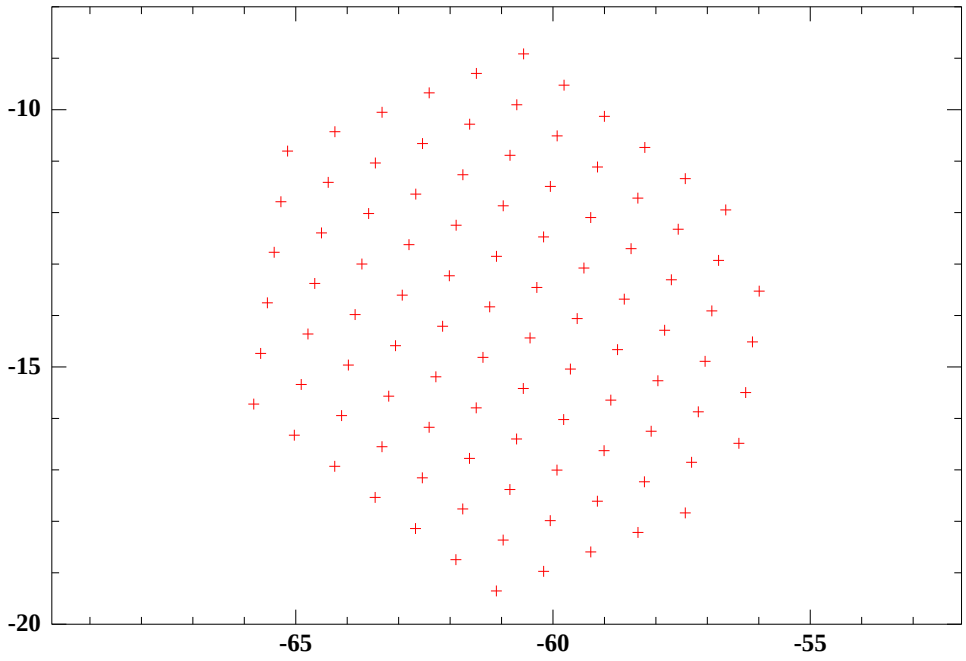


FIG. 8.6 – Solution provided by the GA for $N = 100$ which coincides with the reference configuration $\mathcal{C}^{ref}$.

References

[1] M. Abramowitz & I. A. Stegun, *Handbook of mathematical functions*, Wiley Interscience, 1970.

[2] A. Ben Haj Yedder. MyGa : a Genetic Algorithm in Fortran, available at `http ://cermics.enpc.fr/~benhaj/MyGa/`.

[3] B. Hartke. *Global geometry optimization of clusters using a growth strategy optimized by a Genetic Algorithm.* Chem. Phys. Lett. 240, pp 560-565, 1995.

[4] X. Blanc, C. Le Bris, *Periodicity of the infinite-volume ground state of a one-dimensional quantum model,* Nonlinear Analysis, T.M.A. 48 (6), pp 791-803, 2002.

[5] X. Blanc, C. Le Bris, P-L. Lions, *From molecular models to continuum mechanics*, Archive for Rational Mechanics and Analysis, 164 (4) pp 341-381, 2002.

[6] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Numerical Optimization : Theoretical and Practical Aspects.* Springer-Verlag, New York, 2000.

[7] D. M. Deaven, N. Tit, J. R. Morris and K.M. Ho. *Structural optimization of Lennard-Jones clusters by a genetic algorithm.* Chem. Phys. Lett. 256, pp 195-200, 1996.

[8] P. Engel, *Geometric crystallography. An axiomatic introduction to crystallography*, R. Reidel Publishing Company, 1942

[9] C. A. Floudas, P. M. Pardalos, C. S. Adjiman, W. R. Esposito, Z. H. Gumus, S. T. Harding, J. L. Klepeis, C. A. Meyer and C. A. Schweiger *Handbook of Test Problems in Local and Global Optimization*, Kluwer, Dordrecht, 1999.

[10] C. S. Gardner, C. Radin, *The infinite-volume ground state of the Lennard-Jones potential*, J. Stat. Phys., 20 (6), pp 719-724, 1979.

[11] J. Gu, B. Du, *A Multispace Search Algorithm for Molecular Energy Minimization*, in Discrete Mathematics and Theoretical Computer Science : Global Minimization of Nonconvex Energy Functions, 23, pp 65–87, 1995.

[12] R. Heitman, C. Radin, *Ground states for sticky disks*, J. Stat. Phys. 22, pp 281-287, 1980.

[13] M.R. Hoare, *Structure and dynamics of simple microclusters*, Advan. Chem. Phys. 40, p 49, 1979.

[14] J. E. Lennard-Jones, P. A. Taylor, *Some theoretical calculations of the physical properties of certain crystals*, Proc. Roy. Soc. London Series A 106, 1925, p 476.

[15] E. H. Lieb & B. Simon, *The Thomas-Fermi theory of atoms, molecules and solids*, Adv. in Maths., 23, 1977, pp 22-116.

[16] M. Locatelli, F. Schoen, *Fast Global Optimization of Difficult Lennard-Jones Clusters*, Computational Optimization and Applications, 21 (1), pp 55-70, 2002.

[17] C. Maranas, C. Floudas *A global optimization approach for Lennard-Jones microclusters*, J. Chem. Phys., 97 (10), pp 7667-7677, 1992.

[18] The MathWorks, Inc., MATLAB Reference Guide, 1992.

[19] B.R.A Nijboer, W.J. Ventevogel, *On the configuration of systems of interacting particles with minimum potential energy per particle*, Physica 98A, p 274, 1979.

[20] B.R.A Nijboer, W.J. Ventevogel, *On the configuration of systems of interacting particles with minimum potential energy per particle*, Physica 99A, p 569, 1979.

[21] R.G. Parr, W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, 1989.

[22] C. Radin, *Classical ground states in one dimension*, J. Stat. Phys., 35, p 109, 1983.

[23] C. Radin, *Ground states for soft disks,* J. Stat. Phys., 26, p 365, 1981.

[24] C. Radin, *Low temperature and the origin of crystalline symmetry*, Int. J. Mod. Phys. B 1, pp 1157-1191, 1987.

[25] C. Radin, L. S. Schulmann, *Periodicity of classical ground states*, Phys. Rev. Letters, vol 51, n^o 8, pp 621-622, 1983.

[26] M. S. Stave, A. E. De Pristo, *The structure of Ni_N and Pd_N clusters*, J. Chem. Phys. 97, pp 3386-3398, 1997.

[27] W.J. Ventevogel *On the configuration of a one-dimensional system of interacting particles with minimum potential energy per particle*, Physica 92A, p 343, 1978.

[28] D.J. Wales, J.P.K. Doye, *Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 Atoms* J. Phys. Chem. A, 101, pp 5111-5116 1997.

[29] D. J. Wales, L. J. Munro, J. P. K. Doye, *What can calculations employing empirical potentials teach us about bare transition metal clusters ?*, J. Chem. Soc. Dalton Trans. p 611-624, 1996.

[30] G. L. Xue, *Minimum inter-particle distance at global minimizers of Lennard-Jones clusters*, J. Global Optimization 11, pp 83-90, 1997.

Chapitre 9

Mathematical remarks on the Optimized Effective Potential problem

Dans ce chapitre on présente les résultats théoriques sur le problème Optimized Effective Potential (OEP). Dans ce problème on s'intéresse à l'énergie de Hartree-Fock et à sa comparaison à l'énergie d'un problème approché (OEP).

Mathematical remarks on the Optimized Effective Potential problem

Adel BEN-HAJ-YEDDER, Eric CANCES, Claude LE BRIS
CERMICS, Ecole Nationale des Ponts et Chaussées,

6 et 8 avenue Blaise Pascal, Cité Descartes Champs-sur-Marne,
77455 Marne-la-Vallée Cedex 2, France
{benhaj,cances,lebris}@cermics.enpc.fr

9.1 Motivation

One of the central issue of computational quantum chemistry (see e.g. [1] for an introduction) is the determination of the electronic ground state of molecular system consisting of K nuclei, of charge z_k, and located at known positions $\bar{x}_k$, $1 \leq k \leq K$. Basically, it consists in finding the state Ψ minimizing

$$\inf\{< H_N\Psi, \Psi > \ / \ \Psi \in L^2_a(\mathbf{R}^{3N}),\ \|\Psi\|_{L^2(\mathbb{R}^{3N})} = 1\} \tag{9.1}$$

where the Hamiltonian is given by

$$H_N = -\sum_{i=1}^{N} \Delta_{x_i} + \sum_{i=1}^{N}\left(\sum_{k=1}^{K} \frac{z_k}{|x_i - \bar{x}_k|}\right) + \sum_{1\leq i<j\leq N} \frac{1}{|x_i - x_j|} \tag{9.2}$$

and acts on the position x_i of each of the N electrons. In (9.1), the minimization runs over all antisymmetric functions of $3N$ variables (thus the subscript a). For simplicity, the spin variable is not accounted for. Due to the large size of $L^2_a(\mathbf{R}^{3N})$ for physically relevant values of N, it is not possible to directly attack problem (9.1) and the common practice is to make use of approximations of this problem. One of the most commonly used approximations is the Hartree-Fock approximation (obtained by restricting the minimization in (9.1) to Ψ that are normalized determinants of N functions) and reads :

$$I^{HF} = \inf\{E^{HF}(\phi_1, \ldots, \phi_N), \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\ ,\ 1 \leq i,j \leq N\ ,\ \phi_i \in H^1(\mathbb{R}^3)\} \tag{9.3}$$

where

$$\begin{aligned} E^{HF}(\phi_1, \ldots, \phi_N) \ &= \ \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla\phi_i|^2 - \sum_{i=1}^{N} \int_{\mathbb{R}^3} \left(\sum_{k=1}^{K} \frac{z_k}{|x_i - \bar{x}_k|}\right)|\phi_i|^2 \\ &+ \ \frac{1}{2}\int\int_{(\mathbb{R}^3)^2} \frac{\rho(x)\rho(y)}{|x-y|}dxdy - \frac{1}{2}\int\int_{(\mathbb{R}^3)^2} \frac{|\rho(x,y)|^2}{|x-y|}dxdy, \end{aligned} \tag{9.4}$$

and $\rho(x,y) = \sum_{i=1}^{N} \phi_i(x)\phi_i(y)$, $\rho(x) = \rho(x,x) = \sum_{i=1}^{N} |\phi_i(x)|^2$.

The Hartree-Fock equations are the Euler-Lagrange equations associated to this minimization problem. Up to an orthogonal transform, it can be shown that they read :

$$F_{(\phi_1,...,\phi_N)}\phi_i = -\varepsilon_i\phi_i, \tag{9.5}$$

where the ε_i are real eigenvalues and $F_{(\phi_1,...,\phi_N)}$ is the so-called Fock Hamiltonian

$$F_{(\phi_1,...,\phi_N)} = -\Delta - \sum_{k=1}^{K} \frac{z_k}{|x-\bar{x}_k|} + (\sum_{j\neq i} |\phi_j|^2 \star \frac{1}{|x|}) - (\sum_{j\neq i} \phi_j \bullet \star \frac{1}{|x|})\phi_i. \tag{9.6}$$

As such, equation (9.5) appears as a nonlinear eigenvalue problem involving the Fock operator $F_{(\phi_1,...,\phi_N)}$, which is nonlocal, because of the last term in (9.6) . It is easily understandable that, from the computational viewpoint, constructing the Fock Hamiltonian in a given basis of discretization for the ϕ_i is a costly procedure, in particular because of the nonlocal nature of this operator. As early as in the 1960s (see [7]), the idea has therefore emerged to ask whether equations (9.5) could be rewritten as (or at least approximated by) a system of *local* equations

$$(-\Delta + W)\phi_i = \lambda_i\phi_i, \ i = 1,\dots,N \tag{9.7}$$

for some eigenvalues λ_i and for some *multiplicative* potential W (independent of the index i, but of course possibly dependent of the whole family $(\phi_1,...,\phi_N)$), in a suitable class of regularity (say at least locally integrable). Consequently, the following minimization problem was introduced

$$\begin{aligned} &\text{Minimize } E^{HF}(\phi_1,\dots,\phi_N), \text{ over the set of functions } \phi_i \text{ that satisfy} \\ &\text{the orthonormality constraints of the standard HF problem (9.3) and} \\ &\text{in addition that are eigenfunctions of some operator } -\Delta + W \end{aligned} \tag{9.8}$$

and labelled as the *optimized effective potential* problem (henceforth abbreviated in *OEP problem*). This is to be understood in the sense that one wishes to find the best potential W so that the energy given by some of its eigenfunctions approaches the infimum (9.3).

Let us at once point out that we formulate this problem somewhat vaguely here, for the main concern of the present work will be to give a rigorous mathematical meaning to the formal definition (9.8).

It turns out that the question asked above, that was primarily motivated by considerations on the computational cost, is indeed related to some theoretical questions from quantum chemistry dealing with an alternative theory allowing for a simplification of the original problem (9.1), namely the Density Functional Theory (see e.g. [1,2]). Indeed, a better comprehension of the optimized effective potential problem would give some insight on the construction of accurate exchange-correlation potential for Kohn-Sham models (see [3,4]).

As announced, we intend to give here a possible rigourous foundation to the optimized effective potential problem. As will be seen shortly, our work is a first step, for only very simple cases, sometimes somewhat academic, are addressed. We however believe it provides the main mathematical arguments and may open the way to more thoroughfull studies.

9.2 Setting of the problem and main results

Let us at once make precise that we shall not address the problem of giving a sense to (9.8) in the most general context, but that we shall make three simplifying assumptions.

First, we shall consider *spinless wavefunctions*, as in the above introduction. This simplification is not in fact a limitation, for all the results below can be straigthforwardly extended to the models accounting for spin which are used in computional chemistry, like for instance the restricted Hartree-Fock (RHF) model. It is also to be remarked that for the sake of simplicity, we have chosen to mainly deal with real valued functions. When the consideration of complex valued wavefunctions slightly modify the arguments, we shall indicate it (see in particular Corollary 9.3.2).

A second simplification we shall make, again for the sake of simplicity, is that we shall only consider a molecular system containing *only two electrons*. The consideration of $N > 2$ electrons does not bring any new qualitative phenomenon, but requires rather tedious details that we prefer to avoid. Here and there, we shall however make some remarks in connection with the $N > 2$ case.

Contrarily to the first two, the third simplification we shall make is really restrictive from the mathematical viewpoint. In order to establish some of our main results, we shall restrict our attention to an atom, which means that there is only one nucleus of charge Z, located at $\bar{x} = 0$ (and consequently that $\dfrac{Z}{|x|}$ replaces $\displaystyle\sum_{k=1}^{K} \frac{z_k}{|x - \bar{x}_k|}$ in the energy functional and in the Euler-Lagrange equation), and we shall consider *radially symmetric wavefunctions*. This assumption is restrictive both as results and arguments are concerned. Indeed, spectral theory will play a crucial role in some of our arguments, and it is a well known fact that spectral theory in one dimension (as for radially symmetric functions) features very specific behaviours, in comparison with the situation in dimensions greater than or equal to 2. Likewise, our arguments based upon tools of functional analysis will make an extensive use of the fact that we work in a one dimensional setting. For these reasons, any generalization of our results to the non radial case is to be taken cautiously. In some situations (which is the case of Theorems 9.4.1 and 9.4.2 below, and also for Section 9.5), the same proof and result apply to the general case where functions are not assumed radially symmetric. On the other hand, for some other results, the situation is radically different, as suggested by Proposition 9.3.3 where we give an instance of such a difference with respect to the radial case (Theorem 9.3.1). Let us mention that the optimized effective potential idea has first arised in a radially symmetric setting [7], and that the consideration of this radial case is already quite relevant from the standpoint of

applications in theoretical chemistry.

Let us now define in detail the objects we shall manipulate throughout this article. We have already defined the Hartree-Fock minimization problem (9.3), and the Hartree-Fock energy functional (9.4) in the case of N electrons and K nuclei. For clarity, let us restate them in the restricted case of an atom ($K = 1$) with $N = 2$ electrons :

$$I^{HF} = \inf\{E^{HF}(\phi_1, \phi_2), \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \ , \ 1 \leq i, j \leq 2 \ , \ \phi_i \in H^1(\mathbb{R}^3)\} \tag{9.9}$$

where

$$\begin{aligned} E^{HF}(\phi_1, \phi_2) &= \int_{\mathbb{R}^3} |\nabla \phi_1|^2 + \int_{\mathbb{R}^3} |\nabla \phi_2|^2 - \int_{\mathbb{R}^3} \frac{Z}{|x|} \phi_1^2 - \int_{\mathbb{R}^3} \frac{Z}{|x|} \phi_2^2 \\ &+ \int\int_{(\mathbb{R}^3)^2} \frac{\phi_1^2(x)\phi_2^2(y)}{|x-y|} dxdy - \int\int_{(\mathbb{R}^3)^2} \frac{\phi_1(x)\phi_1(y)\phi_2(x)\phi_2(y)}{|x-y|} dxdy. \end{aligned} \tag{9.10}$$

As announced, we shall mainly restrict ourselves to the case when the functions are assumed to be radially symmetric and therefore to

$$I_r^{HF} = \inf\{E^{HF}(\phi_1, \phi_2), \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \ , \ 1 \leq i, j \leq 2 \ , \ \phi_i \in H_r^1(\mathbb{R}^3)\} \tag{9.11}$$

where $H_r^1(\mathbb{R}^3)$ denotes the set of radially symmetric functions of $H^1(\mathbb{R}^3)$. Accordingly, we shall say that (ϕ_1, ϕ_2) is a solution of the Hartree-Fock equation whenever it satisfies

$$\left\{ \begin{array}{l} -\Delta\phi_1 - \frac{Z}{|x|}\phi_1 + (\phi_2^2 \star \frac{1}{|x|})\phi_1 - (\phi_1\phi_2 \star \frac{1}{|x|})\phi_2 = -\varepsilon_1\phi_1, \\ -\Delta\phi_2 - \frac{Z}{|x|}\phi_2 + (\phi_1^2 \star \frac{1}{|x|})\phi_2 - (\phi_1\phi_2 \star \frac{1}{|x|})\phi_1 = -\varepsilon_2\phi_2, \\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, 1 \leq i, j \leq 2 \end{array} \right. \tag{9.12}$$

Most of the time, ϕ_1 and ϕ_2 will be radially symmetric.

Let us also briefly mention the complex valued case, where the Hartree-Fock minimization problem (possibly for radially symmetric functions, then indicated by the subscript r) reads

$$I_{(r)}^{HF,\mathbb{C}} = \inf\{E^{HF,\mathbb{C}}(\phi_1, \phi_2), \int_{\mathbb{R}^3} \phi_i \phi_j^* = \delta_{ij} \ , \ 1 \leq i, j \leq 2 \ , \ \phi_i \in H_{(r)}^1(\mathbb{R}^3, \mathbb{C})\} \tag{9.13}$$

$$\begin{aligned} E^{HF,\mathbb{C}}(\phi_1,\phi_2) &= \int_{\mathbb{R}^3}|\nabla\phi_1|^2+\int_{\mathbb{R}^3}|\nabla\phi_2|^2-\int_{\mathbb{R}^3}\frac{Z}{|x|}|\phi_1|^2-\int_{\mathbb{R}^3}\frac{Z}{|x|}|\phi_2|^2 \\ &+ \int\int_{(\mathbb{R}^3)^2}\frac{|\phi_1|^2(x)|\phi_2|^2(y)}{|x-y|}dxdy \\ &- \int\int_{(\mathbb{R}^3)^2}\frac{\phi_1(x)\phi_1^*(y)\phi_2^*(x)\phi_2(y)}{|x-y|}dxdy. \end{aligned} \tag{9.14}$$

while the Hartree-Fock equations are

$$\begin{cases} -\Delta\phi_1-\dfrac{Z}{|x|}\phi_1+(|\phi_2|^2\star\dfrac{1}{|x|})\phi_1-(\phi_1\phi_2^*\star\dfrac{1}{|x|})\phi_2=-\varepsilon_1\phi_1, \\ -\Delta\phi_2-\dfrac{Z}{|x|}\phi_2+(|\phi_1|^2\star\dfrac{1}{|x|})\phi_2-(\phi_1^*\phi_2\star\dfrac{1}{|x|})\phi_1=-\varepsilon_2\phi_2, \\ \displaystyle\int_{\mathbb{R}^3}\phi_i\phi_j^*=\delta_{ij}, 1\le i,j\le 2. \end{cases} \tag{9.15}$$

Notation We shall make use of the notation, usual in this context,

$$D(f,g)=\int\int_{(\mathbb{R}^3)^2}\frac{f(x)g(y)}{|x-y|}dxdy, \tag{9.16}$$

whenever this integral makes sense.

9.2.1 Definition of the OEP problems

We now wish to suggest a mathematical definition for the optimized effective potential problem vaguely defined in (9.8). But before we get to this, we would like to introduce a variant of (9.8), namely

> Minimize $E^{HF}(\phi_1,\ldots,\phi_N)$, over the set of functions ϕ_i that satisfy the orthonormality constraints of the standard HF problem (9.3) and in addition that are **the first** N eigenfunctions of some operator $-\Delta+W$. (9.17)

The reason why we introduce such a variant is the following. By a result proven in [5], any N-tuple $(\phi_1,...,\phi_N)$ minimizing the Hartree-Fock energy is a solution of (9.5) that enjoys the following property : the ϕ_i are the first N eigenfunctions of the operator $F_{\phi_1,\ldots,\phi_N}$. Therefore, both for computational reasons (because searching for the first N eigenvalues of a matrix is a specific problem) and for theoretical purposes, it is natural to introduce the variant (9.17). In fact, we shall concentrate most of our attention to this variant, which is indeed the physically relevant version of the OEP problem, and only consider (9.8) as a pedagogic step.

In order to give a sense to (9.8) or respectively (9.17), a major obstacle needs to be overcome. As such, the problem of minimizing upon W is ill-posed, because one lacks of a control on the minimizing sequence W_n in any natural norm. Of course, one could introduce a penalized formulation of the problem, and we will indeed do so in Section 9.5 below, but we prefer to concentrate our efforts on another track. We shall introduce a "weak" formulation of the problems (see (9.20) and (9.26) below), that can be shown to lead to a well posed mathematical problem, and then check, at least formally, that this weak version indeed allows to recover the problem in a strong sense. Let us now motivate our choice for such a weak formulation.

Considering two eigenfunctions ϕ_1 and ϕ_2 of a given operator $-\Delta + W$

$$\left\{ \begin{array}{l} -\Delta\phi_1 + W\phi_1 = \lambda_1\phi_1, \\ -\Delta\phi_2 + W\phi_2 = \lambda_2\phi_2, \end{array} \right. \tag{9.18}$$

it is immediate to see that the following condition, henceforth designated as the *commutation condition*, is fulfilled

$$\phi_2\Delta\phi_1 - \phi_1\Delta\phi_2 = c\phi_1\phi_2 \tag{9.19}$$

with $c = \lambda_2 - \lambda_1$. Conversely, if two functions ϕ_1 and ϕ_2 satisfy (9.19), then they formally are eigenfunctions of $-\Delta + W$ for $W = \dfrac{\Delta\phi_1}{\phi_1}$ respectively for the eigenvalues 0 and c. Thus the idea is to introduce the following minimization problem

$$\begin{array}{rcl} \widetilde{I^{OEP}} & = & \inf\{E^{HF}(\phi_1,\phi_2), \displaystyle\int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\ ,\ 1 \le i,j \le 2, \\ & & \phi_i \in H^1(\mathbb{R}^3), \ \text{ such that for some } c \in \mathbb{R} \\ & & \phi_2\Delta\phi_1 - \phi_1\Delta\phi_2 = c\phi_1\phi_2 \text{ in the sense of } \mathcal{D}'(\mathbb{R}^3)\}, \end{array} \tag{9.20}$$

in order to give a proper meaning to (9.8) in the case of two functions. Of course, an analogous definition can be set, in an obvious way, for $\widetilde{I_r^{OEP}}$ (radial case). Likewise, introducing the two conditions

$$\left\{ \begin{array}{l} \phi_2\Delta\phi_1 - \phi_1\Delta\phi_2 = c\phi_1\phi_2 \\ \phi_2\Delta\phi_1^* - \phi_1^*\Delta\phi_2 = c\phi_1^*\phi_2 \end{array} \right. \tag{9.21}$$

still for c real, one may define $\widetilde{I^{OEP,\mathbb{C}}}$ (complex valued case), and $\widetilde{I_r^{OEP,\mathbb{C}}}$ (radial complex valued case). In the complex case indeed, two commutation conditions are needed to ensure that the potential W formally defined by $W = \dfrac{\Delta\phi_1}{\phi_1}$ is real valued.

The extension of these definitions to the case of N one-electron wavefunctions $(\phi_1, \cdots, \phi_N)$ with $(N-1)$ conditions of the type (9.19)

$$\phi_k\Delta\phi_1 - \phi_1\Delta\phi_k = c_k\phi_1\phi_k \tag{9.22}$$

for $2 \leq k \leq N$ is left to the reader.

One purpose of the present work will be to study the well-posedness of problem (9.20) and show it provides a sound mathematical foundation for the vaguely stated problem (9.8).

In order to now account for the additional condition of being the *first* N eigenfunctions as stated in (9.17), we now go one step further. Suppose we have at hand the first eigenfunction ϕ_1 (with eigenvalue λ_1) and one of the second eigenfunctions ϕ_2 (with eigenvalue λ_2) of some $-\Delta + W$, the two of them forming an orthonormal system. Of course, condition (9.19) is indeed satisfied with some $c = \lambda_2 - \lambda_1 \geq 0$, but we can also assert that

$$\forall \psi \in \mathcal{D}(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \phi_1^2 \, |\nabla \psi|^2 \geq c \left(\int_{\mathbb{R}^3} \psi^2 \, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi \phi_1^2 \big)^2 \right), \tag{9.23}$$

for the same c. Indeed, a simple computation shows that

$$\int_{\mathbb{R}^3} |\nabla(\psi\phi_1)|^2 + \int_{\mathbb{R}^3} (W - \lambda_1) \, (\psi\phi_1)^2 = \int_{\mathbb{R}^3} \phi_1^2 \, |\nabla\psi|^2,$$

thus (9.23) amounts to

$$\big((-\Delta + W)\theta, \theta\big) - \lambda_1 \int_{\mathbb{R}^3} \theta^2 \geq c \left(\int_{\mathbb{R}^3} \theta^2 - \big(\int_{\mathbb{R}^3} \theta\phi_1 \big)^2 \right) \tag{9.24}$$

with $c = \lambda_2 - \lambda_1$, for any function θ which writes $\theta = \psi\phi_1$. Inequality (9.24) obviously holds true, in fact for general θ, because ϕ_1 and ϕ_2 are respectively the first and a second eigenfunction of $-\Delta + W$. In addition, property (9.24) caracterizes ϕ_1 and ϕ_2 as such, among all eigenfunctions of the operator $-\Delta + W$. Indeed, suppose we are given two eigenfunctions ϕ_i and ϕ_j of $-\Delta + W$ such that, according to $c \geq 0$, $\lambda_j - \lambda_i \geq 0$, and such that

$$\big((-\Delta + W)\theta, \theta\big) - \lambda_i \int_{\mathbb{R}^3} \theta^2 \geq (\lambda_j - \lambda_i) \left(\int_{\mathbb{R}^3} \theta^2 - \big(\int_{\mathbb{R}^3} \theta\phi_i \big)^2 \right). \tag{9.25}$$

Then, (formally) testing this condition on $\theta = \phi_1$, the first normalized eigenfunction of $-\Delta + W$, we obtain

$$0 \geq \lambda_1 - \lambda_i \geq (\lambda_j - \lambda_i) \left(1 - \big(\int_{\mathbb{R}^3} \phi_1\phi_i \big)^2 \right).$$

Therefore, either $\big| \int_{\mathbb{R}^3} \phi_1\phi_i \big| \geq 1$, or $\lambda_j = \lambda_i = \lambda_i$, both conditions implying that ϕ_i is the first eigenfunction ϕ_1 (up to a sign), and $\lambda_i = \lambda_1$. Next, testing condition (9.25) (again formally) on any eigenfunction ϕ_k of $-\Delta + W$ different from, thus orthogonal to, ϕ_1, we obtain

$$\lambda_k - \lambda_1 \geq \lambda_j - \lambda_1,$$

which asserts that λ_j is the second eigenvalue λ_2, and that ϕ_j is a second eigenfunction.

Conversely, consider functions ϕ_1 and ϕ_2 such that (9.19) holds. As previously shown, they are formally eigenfunctions of some operator $-\Delta + W$ with $W = \dfrac{\Delta\phi_1}{\phi_1}$. The condition $c \geq 0$ tells that ϕ_2 is associated to an eigenvalue c, greater than (or equal to) the eigenvalue 0 associated to ϕ_1. If in addition (9.23) is satisfied, then it can be written in the same manner as (9.24) (with $\lambda_1 = 0$), and the same formal argument as above shows that ϕ_1 and ϕ_2 are the first two eigenfunctions of the operator.

Of course, all the previous arguments are not rigorous, for in many ocasions we would need to give a proper meaning to the division by ϕ_1. Nevertheless, (9.19), together with $c \geq 0$ and (9.23), appears as a "weak" formulation for the property of being the first two eigenfunctions of some $-\Delta + W$. This consequently justifies the introduction of the problem

$$\begin{aligned}
\widetilde{J^{OEP}} \;=\; & \inf\Big\{ E^{HF}(\phi_1,\phi_2),\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\ ,\ 1 \leq i,j \leq 2, \\
& \phi_i \in H^1(\mathbb{R}^3),\ \text{ such that, for some } c \geq 0 \in \mathbb{R}, \\
& \phi_2\Delta\phi_1 - \phi_1\Delta\phi_2 = c\phi_1\phi_2 \text{ in the sense of } \mathcal{D}'(\mathbb{R}^3), \\
& \text{and such that} \\
& \forall\psi \in \mathcal{D}(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \phi_1^2\,|\nabla\psi|^2 \geq c\left(\int_{\mathbb{R}^3} \psi^2\,\phi_1^2 - \Big(\int_{\mathbb{R}^3} \psi\phi_1^2\Big)^2\right)\Big\} \quad (9.26)
\end{aligned}$$

as a mathematical formulation of (9.17). Of course, an analogous definition can be set, again in an obvious way, for $\widetilde{J_r^{OEP}}$, $\widetilde{J^{OEP,\mathbb{C}}}$, and $\widetilde{J_r^{OEP,\mathbb{C}}}$. Likewise, the definition of problem (9.26) can be extended to the case of N wavefunctions using the $(N-1)$ conditions (9.22) together with the $(N-1)$ inequalities

$$\forall\psi \in \mathcal{D}(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \phi_k^2\,|\nabla\psi|^2 \geq (c_{k+1} - c_k)\left(\int_{\mathbb{R}^3} \psi^2\,\phi_k^2 - \sum_{l=1}^{k}\Big(\int_{\mathbb{R}^3} \psi\phi_l^2\Big)^2\right)$$

for $1 \leq k \leq N-1$, with $c_1 = 0$.

Remark 9.2.1 *It might be useful to remark, and we shall indeed make use of this observation is some of our arguments, that condition (9.23) indeed enforces ϕ_1 to satisfy $\phi_1 \equiv 0$ or $\displaystyle\int_{\mathbb{R}^3} \phi_1^2 = 1$ as soon as $c > 0$. This can indeed easily be proven, letting ψ go to the constant function 1 over $\mathbb{R}^3$.*

We shall study to what extent problem (9.26) provides a rigorous setting for problem (9.17).

9.2.2 Main results

We briefly overview here the main results obtained in the present work. We only give formal statements, postponing the precise statements until the next sections.

First, we investigate the question : can a critical point for the Hartree-Fock energy be a solution to the OEP problem ?

The answer is as follows (Theorem 9.3.1) : in the radial setting, a solution of the Hartree-Fock equations cannot satisfy a condition of the type $\phi_2\Delta\phi_1 - \phi_1\Delta\phi_2 = c\phi_1\phi_2$. The results holds for both real and complex valued functions. Nevertheless, the situation is radically different when allowing for non radially symmetric functions, as shown in Proposition 9.3.3.

Secondly, we show that the $\widetilde{\text{OEP}}$ problems as defined above are well-posed, i.e. that the infimum is attained. This is the purpose of Theorems 9.4.1 and 9.4.2, and their corollaries. There, the wavefunctions are not restricted to be radially symmetric, and may be either real valued or complex valued. For the sake of consistency, we also indicate, in Section 9.5, that penalized forms of the original OEP problems can be considered and show them to be well posed.

We finally explain in Section 9.6 to what extent a minimizer of the problem (9.26) is solution to the original OEP problems as vaguely defined in (9.17). Here, we need to restrict ourselves to radially symmetric functions.

The remainder of this article is devoted to the detailed proofs of the above statements.

9.3 Exploring the link between the HF and the OEP problem

First we shall prove :

Theorem 9.3.1 (Radial case) *A* radial *solution* $(\phi_1, \phi_2) \in \left(H^1_r(\mathbb{R}^3)\right)^2$ *to the Hartree-Fock equations (9.12) cannot satisfy the commutation condition (9.19). A fortiori, it cannot be a solution to (9.7). As a corollary, no radial minimizer of the Hartree-Fock problem is a solution to a system of type (9.7).*

Corollary 9.3.2 *The conclusions of Theorem 9.3.1 hold true* mutatis mutandis *in the case of complex valued functions.*

In Theorem 9.3.1 and its corollary, it is crucial that the functions are radial as shown in the following :

Proposition 9.3.3 (Non radial case) *We may find a pair* (ϕ_1, ϕ_2) *(of non radially symmetric functions) solution to both the Hartree-Fock equations (9.12) and a system of type (9.7).*

We begin by proving Theorem 9.3.1, next show how it can be extended to cover the complex-valued case as claimed in Corollary 9.3.2 above, and then turn to the existence of the counterexample announced of Proposition 9.3.3.

Proof of Theorem 9.3.1 For brevity, we rewrite equations (9.12) in the form :

$$\left\{ \begin{array}{l} -\Delta\phi_1 + V_2\phi_1 - V\phi_2 = 0 \\ -\Delta\phi_2 + V_1\phi_2 - V\phi_1 = 0 \end{array} \right. \tag{9.27}$$

where we have denoted by

$$\begin{aligned} V_1 &= -\frac{Z}{|x|} + \phi_1^2 \star \frac{1}{|x|} + \varepsilon_2, \\ V_2 &= -\frac{Z}{|x|} + \phi_2^2 \star \frac{1}{|x|} + \varepsilon_1, \\ V &= (\phi_1\phi_2) \star \frac{1}{|x|}. \end{aligned}$$

Let us argue by contradiction and assume (ϕ_1, ϕ_2) is an orthonormal system, solution to the above equations (9.27), that in addition satisfies the commutation condition (9.19). By a standard elliptic regularity result, we know that ϕ_1 and ϕ_2 are H^2, continuous on $\mathbb{R}^3$, and that they both are C^∞ outside the origin. In particular, equations (9.27) holds almost everywhere in $\mathbb{R}^3$ and continuously outside the origin. The same applies to (9.19).

Step 1 We begin by showing there exists some open set Ω in $\mathbb{R}^3$ such that, for any $x \in \Omega$

$$\left\{ \begin{array}{l} \left(\phi_1\phi_2 \star \frac{1}{|x|}\right)(x) \neq 0 \\ \phi_1(x)\phi_2(x) \neq 0 \end{array} \right. \tag{9.28}$$

For this purpose, we argue by contradiction, and assume (in view of the continuity) that we have

$$(\phi_1\phi_2 \star \frac{1}{|x|})\phi_1\phi_2 \equiv 0 \,\text{on}\, \mathbb{R}^3. \tag{9.29}$$

If in addition $\phi_1\phi_2 \not\equiv 0$ on $\mathbb{R}^3$, we may find some open set $\omega \neq \emptyset$ such that $\phi_1\phi_2$ has no zero on ω, and thus (9.29) yields $\phi_1\phi_2 \star \frac{1}{|x|} = 0$ on ω. But this implies $\Delta(\phi_1\phi_2 \star \frac{1}{|x|}) = 4\pi\phi_1\phi_2 = 0$ on ω and we reach a contradiction. Therefore (9.29) indeed implies :

$$\phi_1\phi_2 \equiv 0 \,\text{on}\, \mathbb{R}^3. \tag{9.30}$$

Consequently $V = \phi_1\phi_2 \star \frac{1}{|x|} = 0$ and (9.27) reads

$$\left\{ \begin{array}{l} -\Delta\phi_1 + V_2\phi_1 = 0 \\ -\Delta\phi_2 + V_1\phi_2 = 0 \end{array} \right.$$

In addition, since $\phi_1 \not\equiv 0$ because $\int_{\mathbb{R}^3} \phi_1^2 = 1$, we know using (9.30) that $\phi_2 = 0$ on some (non empty) open set. Since ϕ_2 satisfies $-\Delta\phi_2 + V_1\phi_2 = 0$ and vanishes on an open set, we obtain by unique continuation [6] that $\phi_2 \equiv 0$ on $\mathbb{R}^3$. We then reach a contradiction because $\int_{\mathbb{R}^3} \phi_2^2 = 1$, and this concludes this first step.

Step 2 We now show we necessarily have $c = 0$ in equation (9.19) i.e :

$$\phi_1\Delta\phi_2 - \phi_2\Delta\phi_1 = 0. \tag{9.31}$$

Indeed, combining the two equations of (9.27) by multiplying the first one by ϕ_2 and the second one by ϕ_1, next adding the two, we obtain :

$$(-c + V_2 - V_1)\phi_1\phi_2 - V(\phi_2^2 - \phi_1^2) = 0.$$

As we have $\frac{1}{4\pi}\Delta(V_2 - V_1) = \phi_1^2 - \phi_2^2$ and $-\frac{1}{4\pi}\Delta V = \phi_1\phi_2$, we rewrite this equation as :

$$g\Delta f - f\Delta g = 0 \tag{9.32}$$

where $f = V$ and $g = -c + V_2 - V_1$. So,

$$\mathrm{div}(g\nabla f - f\nabla g) = 0$$

and therefore

$$g\frac{df}{dr} - f\frac{dg}{dr} = \frac{a}{r^2},$$

for some real constant a (note that we explicitely use the fact that we work with radially symmetric functions). As f, $\frac{df}{dr}$, g, $\frac{dg}{dr}$ are bounded (this is a consequence of Cauchy-Schwarz and Hardy inequalities) $\frac{a}{r^2}$ must also be bounded when $r \longrightarrow 0$, which implies $a = 0$. It follows that $g\frac{df}{dr} - f\frac{dg}{dr} = 0$. On an open set Ω where f has no zero, as defined by Step 1, it implies that $\frac{d}{dr}(\frac{g}{f}) = 0$, so $g = bf$ on Ω for some constant b, and therefore $\Delta g = b\Delta f$ which yields $(\frac{\phi_2}{\phi_1})^2 - 1 = b\frac{\phi_2}{\phi_1}$ since ϕ_1 has no zero either on Ω, by Step 1. So on some open subset Ω', connex component of Ω, we have $\phi_2 = \alpha\phi_1$ for some constant α. Inserting this in (9.19) yields $c = 0$ and Step 2 is completed.

Step 3

Let us now consider $x \in \mathbb{R}^3$. We claim we have :

– If $\phi_1(x) \neq 0$, there exists some real α_1 and an open set Ω' containing x such that : $\phi_2 = \alpha_1\phi_1$ on Ω'. If in addition $\alpha_1 \neq 0$, then $V_2 - V_1 = (\alpha_1 - \frac{1}{\alpha_1})V$ on Ω'.

– If $\phi_2(x) \neq 0$, there exists some real α_2 and an open set Ω' containing x such that : $\phi_1 = \alpha_2\phi_2$ on Ω'. If in addition $\alpha_2 \neq 0$, then $V_1 - V_2 = (\alpha_2 - \frac{1}{\alpha_2})V$ on Ω'.

We e.g. treat the case $\phi_1(x) \neq 0$. By continuity, there exists an open set Ω' containing x where ϕ_1 has no zero. Integrating (9.31) and arguing as in Step 2, we first deduce the existence of some real constant a such that $\phi_2\frac{d\phi_1}{dr} - \phi_1\frac{d\phi_2}{dr} = \frac{a}{r^2}$ on the whole space, and secondly obtain $a = 0$. This yields $\phi_2 = \alpha_1\phi_1$ on the connex component of Ω' containing x. If in addition $\alpha_1 \neq 0$, system (9.27) reads

$$\begin{cases} -\Delta\phi_1 + V_2\phi_1 - V\alpha_1\phi_1 = 0, \\ -\Delta\phi_2 + V_1\phi_2 - V\frac{\phi_2}{\alpha_1} = 0, \end{cases} \tag{9.33}$$

on this connex component, and combining these two equations we obtain :

$$(\phi_1\Delta\phi_2 - \phi_2\Delta\phi_1) + (V_2 - V_1 - (\alpha_1 - \frac{1}{\alpha_1})V)\phi_1\phi_2 = 0$$

thus, using (9.31) and the fact that $\phi_1\phi_2$ has no zero on Ω',

$$V_2 - V_1 - (\alpha_1 - \frac{1}{\alpha_1})V = 0.$$

The case $\phi_2(x) \neq 0$ is treated in the same manner. Of course when $\phi_1(x)\phi_2(x) \neq 0$ we have $\alpha_1\alpha_2 \neq 0$ and $\alpha_2 = \frac{1}{\alpha_1}$.

Step 4 Let us introduce the function R defined by :

$$R(x) = \begin{cases} 0 & \text{when} \quad \phi_1(x) = \phi_2(x) = 0 \\ V(x)\frac{\phi_2(x)}{\phi_1(x)} & \text{when} \quad \phi_1(x) \neq 0 \\ V(x)\frac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x) & \text{when} \quad \phi_2(x) \neq 0 \end{cases}$$

We claim that R is well defined and $R \in L^\infty(\mathbb{R}^3)$.

In order to prove that R is well defined, the only fact we have to show is that when $\phi_1(x) \neq 0$ and $\phi_2(x) \neq 0$, the two definitions of $R(x)$ yield the same value. It is a simple consequence of Step 3, since for such x : $V(x)\frac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x) = V(x)\frac{\phi_2(x)}{\phi_1(x)}$.

Let us now prove that $R \in L^\infty(\mathbb{R}^3)$. It suffices to consider the different cases

– if $\phi_1(x) = \phi_2(x) = 0$ then $R(x) = 0$,

– if $\phi_1(x) \neq 0$ and $\phi_2(x) = 0$ then $R(x) = V(x)\dfrac{\phi_2(x)}{\phi_1(x)} = 0$,

– if $\phi_1(x) = 0$ and $\phi_2(x) \neq 0$, using $R(x) = V\dfrac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x)$ we have $R(x) = V_2(x) - V_1(x)$,

– if $\phi_1(x)\phi_2(x) \neq 0$, we can make use of both expressions $R(x) = V\dfrac{\phi_2(x)}{\phi_1(x)}$ and $R(x) = V\dfrac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x)$. Therefore, if $|\dfrac{\phi_2(x)}{\phi_1(x)}| \leq 1$, we use $R(x) = V(x)\dfrac{\phi_2(x)}{\phi_1(x)}$ and obtain $|R(x)| = |V\dfrac{\phi_2(x)}{\phi_1(x)}| \leq |V|$. Alternatively, if if $|\dfrac{\phi_1(x)}{\phi_2(x)}| \leq 1$, we use $R(x) = V(x)\dfrac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x)$, and obtain

$$|R(x)| = |V(x)\frac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x)| \leq |V(x)| + |V_2(x) - V_1(x)|.$$

In either case we have $|R(x)| \leq |V(x)| + |V_2(x) - V_1(x)|$, and since V and $V_2 - V_1$ are in $L^\infty(\mathbb{R}^3)$ we conclude that $R \in L^\infty(\mathbb{R}^3)$.

Step 5

Let us now check that both functions ϕ_1 and ϕ_2 are solutions to

$$(-\Delta + V_2 - R)\phi = 0. \tag{9.34}$$

For this purpose, in view of the regularity of the ϕ_i, we only have to check that this equation holds pointwise for all $x \neq 0$.

To begin with, we remark that if for $x \neq 0$ we have $\phi_i(x) = 0$ (for $i = 1$ or $i = 2$) then $\Delta\phi_i(x) = 0$. Indeed, if $\phi_1(x) = \phi_2(x) = 0$, $\Delta\phi_1(x) = \Delta\phi_2(x) = 0$ using (9.27). If $\phi_1(x) \neq 0$ and $\phi_2(x) = 0$, using (9.31) we get $\phi_1(x)\Delta\phi_2(x) = 0$ so $\Delta\phi_2(x) = 0$. And if $\phi_1(x) = 0$ and $\phi_2(x) \neq 0$, using (9.31) again we get $\phi_2(x)\Delta\phi_1(x) = 0$ so $\Delta\phi_1(x) = 0$.

We are now in position to check (9.34) holds for all $x \neq 0$:

(a) If $\phi_1(x) = \phi_2(x) = 0$, then $\Delta\phi_1(x) = \Delta\phi_1(x) = 0$ thus (9.34) holds.

(b) If $\phi_1(x) \neq 0$ and $\phi_2(x) = 0$, then (9.34) is satified by ϕ_2 at x, and, since the first equation of (9.27) gives $-\Delta\phi_1(x) + V_2(x)\phi_1(x) = 0$ and $R(x) = V(x)\dfrac{\phi_2(x)}{\phi_1(x)} = 0$, we have $-\Delta\phi_1(x) + (V_2(x) - R(x))\phi_1(x) = 0$.

(c) If $\phi_1(x) = 0$ and $\phi_2(x) \neq 0$, then

$$-\Delta\phi_1(x) + (V_1(x) - R(x))\phi_1(x) = 0$$

and as the second equation of (9.27) gives $-\Delta\phi_2(x) + V_1(x)\phi_2(x) = 0$,

$$\begin{aligned} -\Delta\phi_2(x) + (V_2(x) - R(x))\phi_1(x) &= -V_1(x)\phi_2(x) + (V_2(x) - R(x))\phi_2(x) \\ &= -V_1(x)\phi_2(x) + V_2(x)\phi_2(x) \\ &\quad -(V(x)\frac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x))\phi_2(x) \\ &= -V(x)\phi_1(x) = 0 \end{aligned}$$

by using $R(x) = V(x)\dfrac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x)$.

(d) If $\phi_1(x)\phi_2(x)(x) \neq 0$ so using the equation (9.27) we obtain :

$$\begin{aligned} -\Delta\phi_1(x) + (V_2(x) - R(x))\phi_1(x) &= -\Delta\phi_1(x) + V_2(x)\phi_1(x) - V\frac{\phi_2(x)}{\phi_1(x)}\phi_1(x) \\ &= -\Delta\phi_1(x) + V_2(x)\phi_1(x) - V(x)\phi_2(x) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} -\Delta\phi_2(x) + (V_2(x) - R(x))\phi_2(x) &= -\Delta\phi_2(x) + V_2(x)\phi_2(x) \\ &\quad -(V(x)\frac{\phi_1(x)}{\phi_2(x)} + V_2(x) - V_1(x))\phi_2(x) \\ &= -\Delta\phi_2(x) + V_1(x)\phi_2(x) - V(x)\phi_1(x) \\ &= 0. \end{aligned}$$

Step 6 We now can conclude the proof. Since ϕ_1 is of norm one and continuous, there exists an open set Ω on which ϕ_1 has no zero. Using Step 3, there exists α_1 such that $\phi_2 = \alpha_1\phi_1$ on a subset Ω' of Ω. Next, by Step 5, ϕ_1 and ϕ_2 are solutions to $(-\Delta + V_2 - R)\phi = 0$, so $\alpha_1\phi_1$ and ϕ_2 are solutions to this equation. Therefore, the functions $\alpha_1\phi_1$ and ϕ_2 are solutions to this equation almost everywhere in $\mathbb{R}^3$, and coincide on Ω'. Hence, $\phi_2 = \alpha_1\phi_1$ everywhere by unique continuation. We reach a contradiction because $\displaystyle\int_{\mathbb{R}^3} \phi_1\phi_2 = 0$ and $\displaystyle\int_{\mathbb{R}^3} \phi_2^2 = 1$. ♢

We now turn to the proof in the case of complex valued functions, which requires slight modifications of the above arguments.

Proof of Corollary 9.3.2

In the case of two complex valued functions, the HF equations (9.27) read :

$$\begin{cases} -\Delta\phi_1 + V_2\phi_1 - V\phi_2 = 0 \\ -\Delta\phi_2^* + V_1\phi_2^* - V\phi_1^* = 0 \end{cases} \tag{9.35}$$

where, $V_1 = -\frac{Z}{|x|}+|\phi_1|^2\star\frac{1}{|x|}+\varepsilon_2$, $V_2 = -\frac{Z}{|x|}+|\phi_2|^2\star\frac{1}{|x|}+\varepsilon_1$ and $V = (\phi_1\phi_2^*)\star\frac{1}{|x|}$. In Step 1, equation (9.28) becomes :

$$\begin{cases} \left(\phi_1\phi_2^* \star \frac{1}{|x|}\right)(x) \neq 0 \\ \phi_1(x)\phi_2^*(x) \neq 0 \end{cases} \tag{9.36}$$

and the proof follows the same pattern. As for Steps 3 to 6, there are only minor changes needed and we leave them to the reader. The only modification that is not straightforward lies in Step 2. The purpose of this step is to show the analogous equation to (9.31), namely

$$\phi_2^*\Delta\phi_1 - \phi_1\Delta\phi_2^* = 0. \tag{9.37}$$

Using the same arguments, we obtain

$$\phi_1\phi_2^*(V_2 - V_1 - c) + V(|\phi_1|^2 - |\phi_2|^2) = 0 \tag{9.38}$$

and thus $|\frac{\phi_2}{\phi_1}|^2 - 1 = b\left(\frac{\phi_2}{\phi_1}\right)^*$ on an open set Ω as defined by Step 1, for some $b \in \mathbb{C}$. Defining $z(x) = \left(\frac{\phi_2}{\phi_1}\right)^*(x)$, this condition reads $|z|^2 - 1 = bz$. Contrary to the real valued case where the conclusion was easily reached, we here have to make a different argument, depending on $b \neq 0$ or $b = 0$. The case $b \neq 0$ is the easy one. Indeed, if $b \neq 0$, it is a simple calculation to show that this implies, for some complex number $\alpha \neq 0$, $\phi_2 = \alpha\phi_1$ on a open subset $\Omega' \subset \Omega$, thus $\phi_2^*\Delta\phi_1 - \phi_1\Delta\phi_2^* = \alpha(\phi_1^*\Delta\phi_1 - \phi_1\Delta\phi_1^*)$, and therefore

$$\phi_1^*\Delta\phi_1 - \phi_1\Delta\phi_1^* = c|\phi_1|^2.$$

It follows that $c = 0$ because the left hand side is imaginary while the right hand side is real.

The case $b = 0$ requires more efforts. We then have $|\phi_2|^2 = |\phi_1|^2$ on Ω. Thus, there exists real valued functions f_1, f_2 and ψ such that $\phi_1(r) = e^{if_1(r)}\psi(r)$ and $\phi_2(r) = e^{if_2(r)}\psi(r)$ on Ω. Rewriting the commutation condition

$$\phi_2\Delta\phi_1^* - \phi_1^*\Delta\phi_2 = c\phi_1^*\phi_2$$

in terms of f_1, f_2 and ψ, we obtain :

$$\psi^2(f_1'' - f_2'') + 2i(f_1' + f_2')(\psi'\psi + \frac{\psi^2}{r}) = c\psi^2.$$

Since $c \in \mathbb{R}$, we have

$$(f_1' + f_2')(\psi'\psi + \frac{\psi^2}{r}) = 0 \,\text{on}\, \Omega. \tag{9.39}$$

If there exists an open subset $\Omega' \subset \Omega$ where $\psi'\phi + \frac{\psi^2}{r}$ is not identically zero, then on such an open set $f_1' + f_2' = 0$, then $\phi_2^* = \alpha\phi_1$ and (9.37) follows. So, in order

to conclude, what we have to rule out is the following situation : on any open set such that (9.36) holds, we have $\phi_1(r) = e^{if_1(r)}\psi(r)$, $\phi_2(r) = e^{if_2(r)}\psi(r)$, $\psi(r) = \frac{\alpha}{r}$ for some constant α. If there is no such open set, the proof is completed, so we suppose there is at least one such Ω_1, say an interval $]\lambda, \mu[$, where $\phi_1(r) = e^{if_1(r)}\psi(r)$, $\phi_2(r) = e^{if_2(r)}\psi(r)$, $\psi(r) = \frac{\alpha}{r}$ for some constant α. We now make a connexity argument. Let us introduce $d \in \overline{\mathbb{R}}$ defined by

$$d = \sup\{ \quad y \text{ such that } \forall x \in]\lambda, y[, \\ \phi_1(r) = e^{if_1(r)}\psi(r),\ \phi_2(r) = e^{if_2(r)}\psi(r),\ \psi(r) = \frac{\alpha}{r}\}.$$

We will show that both cases d finite and $d = +\infty$ lead to a contradiction. Suppose d is finite. By continuity of ψ, $\psi(d) = \frac{\alpha}{d}$ so $\phi_1\phi_2^*(d) \neq 0$. In addition, $\left(\phi_1\phi_2^* \star \frac{1}{|x|}\right)(d) = 0$: otherwise there exists, $\eta > 0$ such that on $]d-\eta, d+\eta[$, (9.36) holds, thus we have $\psi' + \frac{\psi}{r} = 0$ identically and this contradicts the definition of d. In addition, d necessarily is an accumulation point of $\{\phi_1\phi_2^* \star \frac{1}{|x|}(r) = 0\}$. Indeed, if it is not, there exists $\eta > 0$ such that on $]d, d+\eta[$, (9.36) holds, and again we may deduce $\psi' + \frac{\psi}{r} = 0$, which contradicts the definition of d. Therefore, $\Delta(\phi_1\phi_2^* \star \frac{1}{|x|})(d) = 0$, i.e. $(\phi_1\phi_2^*)(d) = 0$ which is false. If we now assume $d = +\infty$, this implies $\psi = \frac{\alpha}{r}$ at infinity, which contradicts $\phi_1 \in L^2(\mathbb{R}^3)$. This concludes the proof of Step 2, and thus that of the Corollary. ♢

Proof of Proposition 9.3.3

We present here an example of some (ϕ_1, ϕ_2) both solution of the Hartree-Fock equations (9.12) and of the Optimized Effective Potential equation (9.7) as announced in Proposition 9.3.3. We search for (ϕ_1, ϕ_2) in the form

$$(\phi_1, \phi_2) = (f(r,\theta)\cos(\varphi), f(r,\theta)\sin(\varphi)) \tag{9.40}$$

where (r, θ, φ) are the spherical coordinates and f is a real valued function. The Hartree-Fock equations (9.12) also read

$$\begin{cases} -\Delta\phi_i - \frac{Z}{|x|}\phi_i + \left(\rho \star \frac{1}{|x|}\right)\phi_i - \int_{\mathbb{R}^3} \frac{\rho(x,y)}{|x-y|}\,\phi_i(y)\,dx\,dy = -\epsilon_i\phi_i \\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij} \end{cases} \tag{9.41}$$

with $\rho(x,y) = \sum_{i=1}^{2} \phi_i(x)\phi_i(y)$ and $\rho(x) = \rho(x,x)$.

If (ϕ_1, ϕ_2) is of the form (9.40) with f satisfying the normalization condition $\int_{\mathbb{R}^3} f^2 = 2$, then $\phi_1 = f(r,\theta)\cos(\varphi)$ and $\phi_2 = f(r,\theta)\sin(\varphi)$ automatically satisfy the orthonormality conditions

$$\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij}.$$

Besides, $\rho(x,y) = f(r_x,\theta_x)\, f(r_y,\theta_y)\, \cos(\varphi_x - \varphi_y)$, $\rho(x) = f(r_x,\theta_x)^2$, and therefore $\int_{\mathbb{R}^3} \frac{\rho(x,y)}{|x-y|}\, \phi_1(y)\, dx\, dy$

$$\begin{aligned}
&= \int_0^{+\infty}\int_0^{\pi}\int_0^{2\pi} \frac{f(r_x,\theta_x)\, f(r_y,\theta_y)\, \cos(\varphi_x-\varphi_y)}{\left(r_x^2 + r_y^2 - 2r_x r_y\left(\cos(\theta_x)\cos(\theta_y) + \sin(\theta_x)\sin(\theta_y)\cos(\varphi_x-\varphi_y)\right)\right)^{1/2}} \\
&\qquad \times f(r_y,\theta_y)\cos(\varphi_y)\, \sin(\theta_y)\, dr_y\, d\theta_y\, d\varphi_y \\
&= \int_0^{+\infty}\int_0^{\pi}\int_0^{2\pi} \frac{f(r_x,\theta_x)\, f(r_y,\theta_y)\, \cos(\varphi)}{\left(r_x^2 + r_y^2 - 2r_x r_y\left(\cos(\theta_x)\cos(\theta_y) + \sin(\theta_x)\sin(\theta_y)\cos(\varphi)\right)\right)^{1/2}} \\
&\qquad \times f(r_y,\theta_y)\cos(\varphi_x-\varphi)\, \sin(\theta_y)\, dr_y\, d\theta_y\, d\varphi \\
&= \int_0^{+\infty}\int_0^{\pi}\int_0^{2\pi} \frac{f(r_x,\theta_x)\, f(r_y,\theta_y)\, \cos(\varphi)}{\left(r_x^2 + r_y^2 - 2r_x r_y\left(\cos(\theta_x)\cos(\theta_y) + \sin(\theta_x)\sin(\theta_y)\cos(\varphi)\right)\right)^{1/2}} \\
&\qquad \times f(r_y,\theta_y)\left(\cos(\varphi_x)\cos(\varphi) + \sin(\varphi_x)\sin(\varphi)\right)\, \sin(\theta_y)\, dr_y\, d\theta_y\, d\varphi \\
&= W_0(x)\, \phi_1(x),
\end{aligned}$$

with

$$W_0(x) = \int_0^{+\infty}\int_0^{\pi}\int_0^{2\pi} \frac{f(r_y,\theta_y)^2\, \cos(\varphi)^2\, \times\, \sin(\theta_y)\, dr_y\, d\theta_y\, d\varphi}{\left(r_x^2 + r_y^2 - 2r_x r_y\left(\cos(\theta_x)\cos(\theta_y) + \sin(\theta_x)\sin(\theta_y)\cos(\varphi)\right)\right)^{1/2}};$$

similarly

$$\int_{\mathbb{R}^3} \frac{\rho(x,y)}{|x-y|}\, \phi_2(y)\, dx\, dy = W_0(x)\, \phi_2(x)$$

(with the same W_0). For (ϕ_1, ϕ_2) of the form (9.40), one therefore has

$$-\Delta\phi_i - \frac{Z}{|x|}\phi_i + \left(\rho \star \frac{1}{|x|}\right)\phi_i - \int_{\mathbb{R}^3} \frac{\rho(x,y)}{|x-y|}\, \phi_i(y)\, dx\, dy = -\Delta\phi_i + W\phi_i$$

where W is a local potential. It remains to exhibit a solution (ϕ_1, ϕ_2) to equations (9.41) of the form (9.40). A simple calculation shows that the goal is reached if one can find $f(r,\theta)$ such that

$$\int_{\mathbb{R}^3} |\nabla f|^2 + \int_{\mathbb{R}^3} \frac{f^2}{r^2 \sin^2\theta} < +\infty \tag{9.42}$$

solution to

$$\begin{cases} -\Delta f + \dfrac{1}{r^2\,\sin^2\theta} f - \dfrac{Z}{r} f + \left(\displaystyle\int_{\mathbb{R}^3} G(x,y)\, f(y)^2\, dy \right) f = -\epsilon f \\ \displaystyle\int_{\mathbb{R}^3} f^2 = 2 \end{cases} \tag{9.43}$$

where $G(x,y)$ is the integral kernel

$$G(x,y) = \frac{\sin(\varphi_x - \varphi_y)^2}{|x-y|}.$$

We are going to prove that such a function f can be obtained by solving the variational problem

$$\inf \left\{ E(u), \quad u \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} u^2 \le 2 \right\} \tag{9.44}$$

where

$$E(u) = \int_{\mathbb{R}^3} |\nabla u|^2 - \int_{\mathbb{R}^3} \frac{Z}{r} u^2 + \int_{\mathbb{R}^3} \frac{u^2}{r^2 \sin^2\theta} + \frac{1}{2} \int \int_{(\mathbb{R}^3)^2} G(x,y)\, u^2(x)\, u^2(y)\, dx\, dy.$$

We proceed as follows.

Step 1. We first prove that the infimum of (9.44) is attained. Let us consider a minimizing sequence (u^n); (u_n) being bounded in $H^1(\mathbb{R}^3)$, we can assume that it converges toward $u \in H^1(\mathbb{R}^3)$, weakly in H^1, strongly in L^p_{loc} for $1 \le p < 6$ and almost everywhere. It is then easy to pass to the limit both in the constraint and in the energy to prove that u is a minimizer of (9.44). As $E(|u|) = E(u)$ for any $u \in H^1(\mathbb{R}^3)$, we can assume in addition that $u \ge 0$.

Step 2. Let $\chi \in \mathcal{D}(\mathbb{R}^3)$ supported in the Ball $B_{1/2} = \{x \in \mathbb{R}^3,\ |x| < 1/2\}$ and such that $\displaystyle\int_{\mathbb{R}^3} \chi^2 = 1$. For $\sigma > 0$ and $\tau > 0$, we denote by

$$\chi_{\sigma,\tau}(x) = \tau^{1/2}\, \sigma^{3/2}\, \chi(\sigma x - e_1)$$

where e_1 is the first unit vector of the cartesian coordinates. As $\sin^2 \ge 3/2$ in $\mathrm{Supp}(\chi_{\sigma,\tau})$ and as $0 < G(x,y) \le \frac{1}{|x-y|}$,

$$E(\chi_{\sigma,\tau}) \le \tau\sigma^2 \int_{\mathbb{R}^3} |\nabla\chi|^2 + \frac{2}{3}\tau\sigma^2 \int_{\mathbb{R}^3} \frac{\chi^2}{|x+e_1|^2} - \tau\sigma \int_{\mathbb{R}^3} \frac{Z}{|x+e_1|}\chi^2 + \tau^2\, \sigma\, D(\chi^2, \chi^2).$$

For τ and σ small enough, $\chi_{\sigma,\tau}$ satisfies the constraint $\displaystyle\int_{\mathbb{R}^3} \chi^2_{\sigma,\tau} \le 2$ and $E(\chi_{\sigma,\tau}) < 0$. Therefore, $u \neq 0$ and consequently u satisfies the Euler-Lagrange equation

$$-\Delta u + \frac{u}{r^2 \sin^2\theta} - \frac{Z}{r} u + \left(\int_{\mathbb{R}^3} G(x,y)\, u(y)^2\, dy \right) u = -\epsilon u. \tag{9.45}$$

Step 3. Assume $\displaystyle\int_{\mathbb{R}^3} u^2 < 2$. Then $\epsilon = 0$ in (9.45) and u is a positive eigenvector of the self-adjoint operator on $L^2(\mathbb{R}^3)$ formally defined by

$$A = -\Delta + \frac{1}{r^2 \sin^2\theta} - \frac{Z}{r} + \left(\int_{\mathbb{R}^3} G(x,y)\, u(y)^2\, dy \right)$$

associated with the eigenvalue 0. As $\sigma_{ess}(A) = [0, +\infty[$, u therefore is the ground state of A. But on the other hand, $(A\chi_{\sigma,1}, \chi_{\sigma_1}) < 0$ when σ is small enough. Indeed

$$\begin{aligned}(A\chi_{\sigma,1}, \chi_{\sigma_1}) \;\leq\; & \sigma^2 \int_{\mathbb{R}^3} |\nabla\chi|^2 + \frac{2}{3}\sigma^2 \int_{\mathbb{R}^3} \frac{\chi^2}{|x+e_1|^2} \\ & -\sigma \left[Z \int_{\mathbb{R}^3} \frac{\chi^2(x)}{|x+e_1|}\, dx - \int\int_{(\mathbb{R}^3)^2} \frac{u(y)^2\, \chi(x)^2}{|x+e_1-\sigma y|}\, dx\, dy \right],\end{aligned}$$

and

$$Z \int_{\mathbb{R}^3} \frac{\chi^2(x)}{|x+e_1|}\, dx - \int\int_{(\mathbb{R}^3)^2} \frac{u(y)^2\, \chi(x)^2}{|x+e_1-\sigma y|}\, dx\, dy \underset{\sigma\to 0}{\longrightarrow} \left(Z - \int_{\mathbb{R}^3} u^2 \right) \int_{\mathbb{R}^3} \frac{\chi^2(x)}{|x+e_1|}\, dx > 0$$

since $\displaystyle\int_{\mathbb{R}^3} u^2 < 2 \leq Z$ by assumption. We thus reach a contradiction. Therefore $\displaystyle\int_{\mathbb{R}^3} u^2 = 2$.

Step 4. The function $\rho = u^2$ is solution to

$$\inf \left\{ \widetilde{E}(\rho), \quad \rho \geq 0,\ \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho \leq 2 \right\} \tag{9.46}$$

with $\widetilde{E}(\rho) = E(\sqrt{\rho})$. As $\widetilde{E}$ is strictly convex on the convex set

$$C = \left\{ \rho \geq 0,\ \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho \leq 2 \right\},$$

the solution ρ to (9.46) is unique. Besides, it follows from the definition of the integral kernel $G(x,y)$ that $\widetilde{E}(\mathcal{R}_{\varphi_0}\rho) = \widetilde{E}(\rho)$ for any $\varphi_0 \in \mathbb{R}$, where $\mathcal{R}_{\varphi_0}$ is the rotation operator defined in spherical coordinates by $(\mathcal{R}_{\varphi_0}\rho)(r,\theta,\varphi) = \rho(r,\theta,\varphi-\varphi_0)$. Consequently, the solution $\rho(r,\theta,\varphi)$ to (9.46) in actually independent on the variable φ ($\rho(r,\theta,\varphi) = \rho(r,\theta)$) and therefore, so is $u = \sqrt{\rho}$ since $u > 0$ in $\mathbb{R}^3 \setminus (\mathbb{R}e_3)$ (by Harnack inequality applied to (9.45)). The function $f(r,\theta) = u(r,\theta)$ therefore satisfies the requirements (9.42)-(9.43). ♢

9.4 The OEP problems are well posed

We will study now the energy of the minimization problems presented above. In the case of real valued functions we have :

Theorem 9.4.1 (Radial or non radial case) *For $Z \geq 2$ (neutral atom or positive ion), there exists a minimizer (ϕ_1, ϕ_2) of the minimization problem $\widetilde{I^{OEP}}$ defined by (9.20). The same conclusion holds for the minimization problem $\widetilde{I_r^{OEP}}$ defined analogously and restricted to radially symmetric functions.*

Theorem 9.4.2 (Radial or non radial case) *For $Z \geq 2$ (neutral atom or positive ion), there exists a minimizer (ϕ_1, ϕ_2) of the minimization problem $\widetilde{J^{OEP}}$ defined by (9.26). The same conclusion holds for the minimization problem $\widetilde{J_r^{OEP}}$ defined analogously and restricted to radially symmetric functions.*

Corollary 9.4.3 *The conclusions of Theorems 9.4.1 and 9.4.2 hold true for complex valued functions, ie for the minimization problems $\widetilde{I^{OEP,\mathbb{C}}}$, $\widetilde{I_r^{OEP,\mathbb{C}}}$, $\widetilde{J^{OEP,\mathbb{C}}}$, $\widetilde{J_r^{OEP,\mathbb{C}}}$.*

As we know that there exists a minimizer of I^{HF} problem, using Theorem 9.3.1, we obtain, in the radial case,

Theorem 9.4.4 (Radial case) *For $Z \geq 2$ (neutral atom or positive ion), $I_r^{HF} < I_r^{OEP} \leq J_r^{OEP}$.*

Corollary 9.4.5 *The conclusion of Theorem 9.4.4 holds true in the complex valued case.*

This section is articulated as follows. We first prove Theorem 9.4.1 for general functions (not necessarily radially symmetric), the proof of the radial case being the same. Next, we prove Theorem 9.4.2, again in the general case. The proof of Theorem 9.4.4 is straightforward and we skip it. We also skip the proofs of Corollary 9.4.3 and Corollary 9.4.5, that follow the same lines as those of Theorem 9.4.1 and Theorem 9.4.2 respectively.

Proof of Theorem 9.4.1

Step 1

We begin by proving an *a priori* estimate of the energy :

$$\widetilde{I^{OEP}} < I = \inf \left\{ \int_{\mathbb{R}^3} \left(|\nabla \psi|^2 - \frac{Z}{|x|} \psi^2 \right), \ \int_{\mathbb{R}^3} \psi^2 = 1 \right\} < 0. \tag{9.47}$$

For this purpose, we consider ψ_1 and ψ_2 the first two normalized eigenfunctions of the operator $(-\Delta - \dfrac{Z}{|x|})$ on $L^2(\mathbb{R}^3)$ which are defined by :

$$\psi_1(r,\theta,\varphi) = \left(\frac{Z}{2}\right)^{3/2} \frac{e^{-Zr/2}}{\sqrt{\pi}} \quad \text{and} \quad \psi_2(r,\theta,\varphi) = \left(\frac{Z}{2}\right)^{3/2} \frac{\left(1 - \frac{Zr}{4}\right) e^{-\frac{Zr}{4}}}{\sqrt{8\pi}}. \tag{9.48}$$

Using the notation (9.16), we have :

$$\begin{aligned}
\int_{\mathbb{R}^3} |\nabla\psi_1|^2 &= \frac{Z^2}{4}, \\
\int_{\mathbb{R}^3} |\nabla\psi_2|^2 &= \frac{Z^2}{16}, \\
-\int_{\mathbb{R}^3} \frac{Z}{|x|}\psi_1^2 &= -\frac{Z^2}{2}, \\
-\int_{\mathbb{R}^3} \frac{Z}{|x|}\psi_2^2 &= -\frac{Z^2}{8}, \\
D(\psi_1^2, \psi_2^2) &= \frac{17}{162}Z, \\
D(\psi_1\psi_2, \psi_1\psi_2) &= \frac{8}{729}Z.
\end{aligned} \tag{9.49}$$

Therefore, for any $Z \geq 2$,

$$I = \inf\left\{\int_{\mathbb{R}^3}\left(|\nabla\psi|^2 - \frac{Z}{|x|}\psi^2\right),\ \int_{\mathbb{R}^3}\psi^2 = 1\right\} = \int_{\mathbb{R}^3}\left(|\nabla\psi_1|^2 - \frac{Z}{|x|}\psi_1^2\right) = -\frac{Z^2}{4}$$

and, since (ψ_1, ψ_2) are admissible test functions for $\widetilde{I^{OEP}}$,

$$\widetilde{I^{OEP}} \leq E^{HF}(\psi_1, \psi_2) = Z^2\left(\frac{1}{4} + \frac{1}{16} - \frac{1}{2} - \frac{1}{8}\right) + Z\left(\frac{17}{162} - \frac{16}{729}\right) < -\frac{Z^2}{4} \tag{9.50}$$

for any $Z \geq 2$. Inequality (9.47) follows.

Step 2

Let us now consider a minimizing sequence (ϕ_1^n, ϕ_2^n) of the $\widetilde{EQ:OEP}$ problem (9.20). As this sequence is bounded in $H^1(\mathbb{R}^3)$, we can extract a subsequence that weakly converges in $H^1(\mathbb{R}^3)$ to (ϕ_1, ϕ_2). The weak limit (ϕ_1, ϕ_2) satisfies $E^{HF}(\phi_1, \phi_2) \leq \widetilde{I^{OEP}}$ and $(\int_{\mathbb{R}^3} \phi_i\phi_j) \leq (\delta_{ij}), \quad 1 \leq i, j \leq 2$.

Proving that (ϕ_1, ϕ_2) is a minimizer of (9.20) amounts to proving that (ϕ_1, ϕ_2) also satisfies both conditions

$$\phi_1\Delta\phi_2 - \phi_2\Delta\phi_1 = c\phi_1\phi_2 \ \text{ for some } c \in \mathbb{R} \tag{9.51}$$

$$\int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \quad 1 \leq i, j \leq 2. \tag{9.52}$$

We devote this second step to the proof of (9.51). For each n, we have some real constant c^n such that

$$\phi_1^n\Delta\phi_2^n - \phi_2^n\Delta\phi_1^n = c^n\phi_1^n\phi_2^n. \tag{9.53}$$

We multiply by some arbitrary $\psi \in \mathcal{D}(\mathbb{R}^3)$ and integrate to obtain :

$$\int_{\mathbb{R}^3} (\phi_2^n \nabla \phi_1^n - \phi_1^n \nabla \phi_2^n) \nabla \psi = c^n \int_{\mathbb{R}^3} \phi_1^n \phi_2^n \psi. \tag{9.54}$$

In order to pass to the limit in the left hand side of (9.54), we remark that (ϕ_1^n, ϕ_2^n) weakly converges to (ϕ_1, ϕ_2) in $(H^1(\mathbb{R}^3))^2$, so (ϕ_1^n, ϕ_2^n) strongly converges to (ϕ_1, ϕ_2) in $(L^2_{loc}(\mathbb{R}^3))^2$ and $(\nabla\phi_1^n, \nabla\phi_2^n)$ weakly converges to $(\nabla\phi_1, \nabla\phi_2)$ in $(L^2(\mathbb{R}^3))^2$. Thus $\phi_2^n \nabla \phi_1^n$ and $\phi_1^n \nabla \phi_2^n$ respectively weakly converge to $\phi_2 \nabla \phi_1$ and $\phi_1 \nabla \phi_2$ in $L^1_{loc}(\mathbb{R}^3)$. This allows to pass to the limit in the left hand side.

For the right hand side of (9.54) we proceed as follows. If the real sequence c^n is not bounded, we can extract a subsequence, still denoted by c^n, such that $|c^n| \longrightarrow +\infty$. Then necessarily $\phi_1\phi_2 = 0$, otherwise we may choose $\psi \in \mathcal{D}(\mathbb{R}^3)$ such that $c^n \displaystyle\int_{\mathbb{R}^3} \phi_1^n \phi_2^n \psi \longrightarrow \infty$, and this cannot occur since the left hand side of (9.54) converges. Next, the fact that $\phi_1\phi_2 = 0$ and that $\phi_1\Delta\phi_2 - \phi_2\Delta\phi_1 = 0$ in the sense of $\mathcal{D}'(\mathbb{R}^3)$, and (9.51) is trivially satisfied.

Suppose now that c^n is bounded. Then we can extract a subsequence, still denoted by c^n, that converges to some real constant c, and therefore $c^n\phi_1^n\phi_2^n$ converges in $L^1_{loc}(\mathbb{R}^3)$, say. Equation (9.51) follows. The final two steps are devoted to the proof of the orthonormality condition (9.52).

Step 3

We here prove, that, up to a rotation, we may always assume without loss of generality that

$$\int_{\mathbb{R}^3} \phi_1\phi_2 = 0. \tag{9.55}$$

If the constant c is different from 0 in (9.51) then we have, integrating this equation over the whole space, $c \displaystyle\int_{\mathbb{R}^3} \phi_1\phi_2 = 0$ and (9.55) follows. In order to make this rigorous, since (9.51) only holds in the sense of distributions, say, we introduce a smooth cut-off function χ_R which has value 1 on the ball B_R, 0 outside the ball B_{R+1} and has values in $[0,1]$ on $B_R^c \cap B_{R+1}$, with $\|\chi_R\|_{C^1} \leq 1$. Then we write

$$\begin{aligned} < \phi_2\Delta\phi_1, \chi_R > &= -\int_{\mathbb{R}^3} \chi_R \nabla\phi_1\nabla\phi_2 - \int_{\mathbb{R}^3} \phi_2\nabla\phi_1\nabla\chi_R \\ &= -\int_{B_R} \nabla\phi_1\nabla\phi_2 - \int_{B_R^c\cap B_{R+1}} \big(\chi_R\nabla\phi_1\nabla\phi_2 + \phi_2\nabla\phi_1\nabla\chi_R\big). \end{aligned}$$

As R goes to infinity, the first term goes to $\displaystyle\int_{\mathbb{R}^3} \nabla\phi_1\nabla\phi_2$, while the second one goes to zero using the Cauchy-Schwarz inequality and observing that both ϕ_i are $H^1(\mathbb{R}^3)$ while χ_R is uniformly bounded in C^1.

On the other hand, suppose now $c = 0$ in (9.51). We then replace (ϕ_1, ϕ_2) by $(\widetilde{\phi_1}, \widetilde{\phi_2})$ defined by :

$$\begin{pmatrix} \widetilde{\phi_1} \\ \widetilde{\phi_2} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$$

For any θ, all the following conditions are satisfied

$$\begin{cases} \displaystyle\int_{\mathbb{R}^3} \widetilde{\phi_i}^2 \leq 1, \ 1 \leq i, j \leq 2 \\ \widetilde{\phi_1}\Delta\widetilde{\phi_2} - \widetilde{\phi_2}\Delta\widetilde{\phi_1} = 0 \\ E^{HF}(\widetilde{\phi_1}, \widetilde{\phi_2}) = E^{HF}(\phi_1, \phi_2), \end{cases} \tag{9.56}$$

We then choose θ such that we precisely have

$$\int_{\mathbb{R}^3} \widetilde{\phi_1}\widetilde{\phi_2} = 0.$$

In this manner, all the properties satisfied by (ϕ_1, ϕ_2) are shared by $(\widetilde{\phi_1}, \widetilde{\phi_2})$ with in addition orthogonality. From now on, we forget the notation $\widetilde{\phi_i}$ and simply use ϕ_i, considering that (9.55) is satisfied. We also know that $\displaystyle\int_{\mathbb{R}^3} \phi_i^2 \leq 1$ and there remains now to prove that both ϕ_i are of unit norm.

Step 4

We argue by contradiction and intend to show that

$$\int_{\mathbb{R}^3} \phi_1^2 < 1,$$

say, cannot hold.

For brevity, we denote by

$$D = D(\phi_1^2, \phi_2^2) - D(\phi_1\phi_2, \phi_1\phi_2),$$

(which, we recall, is a nonnegative quantity), and for $i = 1, 2$,

$$A_i = \int_{\mathbb{R}^3} |\nabla\phi_i|^2 - \frac{Z}{|x|}\phi_i^2$$

and, when it makes sense, $\alpha_i = \dfrac{1}{\sqrt{\int_{\mathbb{R}^3} \phi_i^2}}$. In addition, we recall the notation

$$I = \inf\left\{ \int_{\mathbb{R}^3} \left(|\nabla\psi|^2 - \frac{Z}{|x|}\psi^2 \right), \ \int_{\mathbb{R}^3} \psi^2 = 1 \right\}.$$

With the above notations,

$$E^{HF}(\phi_1, \phi_2) = A_1 + A_2 + D,$$

while the orthonormal family $(\alpha_1\phi_1, \alpha_2\phi_2)$ (in view of (9.55)) has energy

$$E^{HF}(\alpha_1\phi_1, \alpha_2\phi_2) = \alpha_1^2 A_1 + \alpha_2^2 A_2 + \alpha_1^2\alpha_2^2 D.$$

To begin with, we rule out the case when one, or both, of the ϕ_i is identically zero. Suppose e.g. $\phi_1 \equiv 0$. Then

$$\widetilde{I^{OEP}} \geq E^{HF}(\phi_1, \phi_2) = A_2 \geq \left(\int_{\mathbb{R}^3} \phi_2^2\right) I \geq I,$$

which contradicts (9.47). We now can suppose that both α_i are well defined. We remark that we have $\alpha_i \geq 1$ and thus, by definition of I, $\alpha_i^2 A_i \geq I$, for $i = 1, 2$.

Suppose we have

$$A_2 + \alpha_1^2 D \geq 0 \text{ or } A_1 + \alpha_2^2 D \geq 0.$$

Then, assuming for instance that the first assertion holds, we have

$$\begin{aligned} E^{HF}(\phi_1, \phi_2) &= A_1 + A_2 + D \\ &\geq A_1 + (1 - \frac{1}{\alpha_1^2})A_2 \\ &\geq \left(\frac{1}{\alpha_1^2} + \frac{1}{\alpha_2^2}(1 - \frac{1}{\alpha_1^2})\right) I \\ &\geq I, \end{aligned}$$

because $I < 0$ and

$$\frac{1}{\alpha_1^2} + \frac{1}{\alpha_2^2}(1 - \frac{1}{\alpha_1^2}) = \frac{\alpha_1^2 + \alpha_2^2 - 1}{\alpha_1^2\alpha_2^2} \leq 1$$

since $\alpha_i \geq 1$. Clearly, since

$$E^{HF}(\phi_1, \phi_2) \leq \widetilde{I^{OEP}}$$

this contradicts (9.47).

On the other hand, suppose we have

$$A_2 + \alpha_1^2 D < 0 \text{ and } A_1 + \alpha_2^2 D < 0.$$

Then

$$\begin{aligned} E^{HF}(\alpha_1\phi_1, \alpha_2\phi_2) &= \alpha_1^2 A_1 + \alpha_2^2(A_2 + \alpha_1^2 D) \\ &< \alpha_1^2 A_1 + A_2 + \alpha_1^2 D \\ &\quad \text{since } \alpha_2 \geq 1 \\ &< A_1 + A_2 + D \\ &\quad \text{since } \alpha_1 > 1, \text{and } A_1 + D \leq A_1 + \alpha_2^2 D < 0, \\ &= E^{HF}(\phi_1, \phi_2) \\ &\leq \widetilde{I^{OEP}} \end{aligned}$$

and again we reach a contradiction because $E^{HF}(\alpha_1\phi_1, \alpha_2\phi_2)$ should be greater than or equal to $\widetilde{I^{OEP}}$ for it satisfies the contraints.

This concludes the proof. ♢

Proof of Theorem 9.4.2

We now indicate the slight modifications that need to be made in the arguments of the proof of Theorem 9.4.1 in order to apply to the minimization problem $\widetilde{J^{OEP}}$.

For Step 1, we only remark that the pair (ψ_1, ψ_2) indeed satisfies all the constraints of problem $\widetilde{J^{OEP}}$, and therefore we have

$$\widetilde{J^{OEP}} < I \ < 0. \tag{9.57}$$

In Step 2, considering a minimizing sequence (ϕ_1^n, ϕ_2^n) for $\widetilde{J^{OEP}}$, we may as above assume it weakly converges in H^1 to some (ϕ_1, ϕ_2) which satisfies $E^{HF}(\phi_1, \phi_2) \leq \widetilde{J^{OEP}}$ and $(\int_{\mathbb{R}^3} \phi_i\phi_j) \leq (\delta_{ij}), \quad 1 \leq i, j \leq 2$. We next pass to the limit in the condition

$$\forall \psi \in \mathcal{D}(\mathbb{R}^3), \ \int_{\mathbb{R}^3} (\phi_1^n)^2 \, |\nabla\psi|^2 \geq c^n \left(\int_{\mathbb{R}^3} \psi^2 \, (\phi_1^n)^2 - \big(\int_{\mathbb{R}^3} \psi (\phi_1^n)^2 \big)^2 \right), \tag{9.58}$$

which will conclude the proof of Step 2. For this purpose, we first simply use the fact that ϕ_1^n strongly converges locally, say in L^2_{loc}. So all integrals in (9.58) converge for ψ fixed. Next, two cases may occur. Either the weak limit ϕ_1 of ϕ_1^n is identically zero, and therefore the condition

$$\forall \psi \in \mathcal{D}(\mathbb{R}^3), \ \int_{\mathbb{R}^3} \phi_1^2 \, |\nabla\psi|^2 \geq c \left(\int_{\mathbb{R}^3} \psi^2 \, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi \phi_1^2 \big)^2 \right). \tag{9.59}$$

is trivially satisfied, for any c. Or, $\phi_1 \not\equiv 0$, and therefore we may find some $\psi \in \mathcal{D}(\mathbb{R}^3)$ such that

$$\int_{\mathbb{R}^3} \psi^2 \, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi \phi_1^2 \big)^2 \neq 0.$$

Therefore, we have

$$\limsup c^n \leq \frac{\displaystyle\int_{\mathbb{R}^3} \phi_1^2 \, |\nabla\psi|^2}{\displaystyle\int_{\mathbb{R}^3} \psi^2 \, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi \phi_1^2 \big)^2},$$

which shows that c^n is a bounded sequence. We thus may assume it converges, to some $c \geq 0$, and pass to the limit in each term of (9.58) to obtain (9.59).

In addition, we may also pass to the limit in the commutation condition to obtain

$$\phi_2 \Delta\phi_1 - \phi_1 \Delta\phi_2 = c\phi_1\phi_2.$$

This concludes Step 2.

For Step 3, we only make the following additionnal comment.

When $c \neq 0$, we have as above, by integration, $\int_{\mathbb{R}^3} \phi_1\phi_2 = 0$, and Step 3 is completed. In the case $c = 0$, condition (9.59) is indeed empty, as the integral of a nonnegative function is always nonnegative. Therefore, (ϕ_1, ϕ_2) may be replaced by $(\widetilde{\phi_1}, \widetilde{\phi_2})$, so that $\displaystyle\int_{\mathbb{R}^3} \widetilde{\phi_1}\widetilde{\phi_2} = 0$, keeping the property that (9.59) is satisfied, again with $c = 0$.

For Step 4, we remark the following. If condition (9.59) is satisfied by ϕ_1 for some $c \geq 0$, then $\alpha_1\phi_1$ also satifies it, whenever $\alpha_1 \geq 1$. Indeed, it suffices to remark that

$$\begin{aligned}
\alpha_1^2 \int_{\mathbb{R}^3} \phi_1^2\, |\nabla\psi|^2 &\geq c\,\alpha_1^2 \left(\int_{\mathbb{R}^3} \psi^2\, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi\phi_1^2 \big)^2 \right) \\
&\geq c \left(\int_{\mathbb{R}^3} \psi^2\, (\alpha_1\phi_1)^2 - \frac{1}{\alpha_1^2} \big(\int_{\mathbb{R}^3} \psi(\alpha_1\phi_1)^2 \big)^2 \right) \\
&\geq c \left(\int_{\mathbb{R}^3} \psi^2\, (\alpha_1\phi_1)^2 - \big(\int_{\mathbb{R}^3} \psi(\alpha_1\phi_1)^2 \big)^2 \right)
\end{aligned}$$

since $\alpha_1 \geq 1$. Therefore we can make use of the same argument as in Step 4 of the proof of Theorem 9.4.1 without modification. ♢

9.5 Penalized form of the OEP problem

The weak forms of the OEP problems (9.8) and (9.17) introduced above in (9.20) and (9.26) may be considered, from a certain point of view, as too *weak*. We shall see that, at least formally and in the simple radial case, they can be shown to be "equivalent" (note the quotes!) to the original problem in the "strong" form. Nevertheless, it remains that from a rigorous viewpoint we are not able to show the equivalence and therefore other tracks for giving a sense (9.8) and (9.17) may be pursued. One of such track is a penalization strategy, where one introduces a control on the potential W in order to be able to pass to the limit in minimizing sequences. From the computational standpoint, such a strategy is not surprising and is efficient in many other settings.

In view of the above motivation, we introduce, for any $\varepsilon > 0$, the following penalized version of problem (9.8)

$$\begin{aligned}
I_\varepsilon^{OEP} \;=\; & \inf\{ E^{HF}(\phi_1,\phi_2) \,+\, \varepsilon \left(\|\mu\|_X + \|V\|_Y \right), \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\ ,\ 1 \leq i,j \leq 2, \\
& \phi_i \in H^1(\mathbb{R}^3),\ \text{ such that for some } \lambda_1,\lambda_2 \in \mathbb{R},\, \mu \in X,\, V \in Y \\
& (-\Delta - \frac{Z}{|x|} + \mu \star \frac{1}{|x|} + V)\, \phi_i = \lambda_i\phi_i \text{ in the sense of } \mathcal{D}'(\mathbb{R}^3) \,\},
\end{aligned} \tag{9.60}$$

In this definition, the functional space X is, for instance, chosen to be L^p for some $1 \leq p < 3/2$ and the functional space Y as L^q for, say, $q = 3/2$. Of course, our choice is arbitrary, and other functional spaces could be chosen, provided they satisfy some technical assumptions that allow for the arguments that will follow in this section. However, we do not want to enter such technicalities, and leave such easy extensions to the reader. Our purpose is only to show that such a penalized problem can be properly stated and solved. The point is that the class of potentials W (according to the notation of (9.8)) that we have chosen, namely $-\dfrac{Z}{|x|} + \mu \star \dfrac{1}{|x|} + V$, with such μ and V, contains some reasonable potentials, relevant from the application viewpoint, so far as we can judge. Of course, such a form is reminiscent of the form of the Fock potential and has of course been mimicked on it.

The same penalization technique can be applied to (9.17) and it leads to the formulation

$$\begin{aligned} J_\varepsilon^{OEP} \;=\; & \inf\{E^{HF}(\phi_1,\phi_2) \,+\, \varepsilon\left(\|\mu\|_X + \|V\|_Y\right), \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\ ,\ 1 \leq i,j \leq 2, \\ & \phi_i \in H^1(\mathbb{R}^3), \quad \text{such that } (\phi_1,\lambda_1)(\text{resp. } (\phi_2,\lambda_2)) \\ & \text{is the first (resp. a second) eigenvector/eigenvalue of the operator} \\ & -\Delta - \frac{Z}{|x|} + \mu \star \frac{1}{|x|} + V \ \text{ for some } \mu \in X,\, V \in Y\}, \end{aligned} \tag{9.61}$$

Like in the previous sections, the minimizations problems $I_{r,\varepsilon}^{OEP}$, $I_\varepsilon^{OEP,\mathbb{C}}$, $I_{r,\varepsilon}^{OEP,\mathbb{C}}$, $J_{r,\varepsilon}^{OEP}$, $J_\varepsilon^{OEP,\mathbb{C}}$, $J_{r,\varepsilon}^{OEP,\mathbb{C}}$, with self-explanatory notations, can be defined accordingly.

In this section, we begin by studying the problem I_ε^{OEP}. For all the other "usual" problems $I_{r,\varepsilon}^{OEP}$, $I_\varepsilon^{OEP,\mathbb{C}}$, $I_{r,\varepsilon}^{OEP,\mathbb{C}}$, the proofs basically follow the same lines and the result of Theorem 9.5.1 below holds *mutatis mutandis*. For brevity, we skip all of them. Next, we study problem J_ε^{OEP}. Our proof can be extended (but we do not do so) to the complex valued case $J_\varepsilon^{OEP,\mathbb{C}}$, and the radial cases $J_{r,\varepsilon}^{OEP}$, $J_{r,\varepsilon}^{OEP,\mathbb{C}}$.

Theorem 9.5.1 *For $Z \geq 2$ (neutral atom or positive ion), the minimization problem (9.60) admits a minimizer.*

Proof of Theorem 9.5.1

The proof mimicks that of Theorem 9.4.1, so we will detail only the differences. *Step 1* consists in showing that

$$I_\varepsilon^{OEP} \,<\, I, \tag{9.62}$$

as defined by the right-hand side of (9.47), and this property is a straightforward consequence of the fact that (ψ_1,ψ_2) defined by (9.48) satifies the constraints of (9.60) (with $\mu = V = 0$), and $E^{HF}(\psi_1,\psi_2) < I$ as shown in (9.50).

Step 2 We consider a minimizing sequence (ϕ_1^n,ϕ_2^n), associated with functions μ^n and V^n respectively in $X = L^p$ and $Y = L^q$. As the Hartree-Fock energy is bounded

from above, we may assume (ϕ_1^n, ϕ_2^n) weakly converges to a limit (ϕ_1, ϕ_2) in $(H^1)^2$. In addition, due to the penalty term, the functions μ^n and V^n may also be assumed to weakly converge in X and Y respectively. In view of the eigenvalue equation

$$(-\Delta - \frac{Z}{|x|} + \mu^n \star \frac{1}{|x|} + V^n)\, \phi_i^n = \lambda_i^n \phi_i^n,$$

the eigenvalues

$$\lambda_i^n = \int_{\mathbb{R}^3} |\nabla \phi_i^n|^2 - \int_{\mathbb{R}^3} \frac{Z}{|x|}(\phi_i^n)^2 + \int_{\mathbb{R}^3} (\mu^n \star \frac{1}{|x|})(\phi_i^n)^2 + \int_{\mathbb{R}^3} V_n (\phi_i^n)^2$$

are also bounded (each of the last three terms of the right-hand side can indeed be treated by Hölder type inequalities, the conditions $1 \leq p < 3/2$ and $q = 3/2$ playing here a role), and therefore may be assumed to also converge, to some λ_i, as n goes to infinity. By Solobev compact imbeddings, we have the local strong convergences of ϕ_i^n in L^r (at least) for $1 \leq r < 6$, and therefore it is then easy to pass to the limit locally in the equations to obtain

$$(-\Delta - \frac{Z}{|x|} + \mu \star \frac{1}{|x|} + V)\, \phi_i = \lambda_i \phi_i. \tag{9.63}$$

As a consequence of the weak convergence in H^1, we have $\displaystyle\int_{\mathbb{R}^3} \phi_i \phi_j \leq 1$ in the sense of symmetric matrices, and there now remains to prove the orthonormality constraint on (ϕ_1, ϕ_2) to conclude the proof.

Step 3 is exactly the same as that in the proof of Theorem 9.4.1, $\lambda_2 - \lambda_1$ playing the role of c, of course.

Step 4 also is in the same vein. Indeed, the only three ingredients that are used are (a) the orthogonality $\displaystyle\int_{\mathbb{R}^3} \phi_1 \phi_2 = 0$ as produced by Step 3, (b) the property that

$$I_\varepsilon^{OEP} \geq E^{HF}(\phi_1, \phi_2) + \varepsilon \left(\|\mu\|_X + \|V\|_Y \right)$$

due to the weak convergences at hand, and (c) the fact that $I_\varepsilon^{OEP} < I$ as remarked in Step 1. Note also that the penalty term, being nonnegative and independent of the norm of ϕ_i does not perturbate the various inequalities involved in the argument. This therefore concludes the proof. ♢

Remark 9.5.2 *It is unfortunately not known whether*

$$\lim_{\varepsilon \longrightarrow 0} I_\varepsilon^{OEP} = \widetilde{I^{OEP}}. \tag{9.64}$$

We now turn to the proof of

Theorem 9.5.3 *For $Z \geq 2$ (neutral atom or positive ion), the minimization problem J_ε^{OEP} defined by (9.61) admits a minimizer.*

Proof of Theorem 9.5.3

We again refer to the usual 4 steps, like in the previous proof. Step 1 is of course unchanged as (ψ_1, ψ_2) are the first two eigenfunctions of $-\Delta - \dfrac{Z}{|x|}$. Next, we need to make some slight modifications of the argument. Consider a miminizing sequence denoted as above. Since ϕ_1^n is the first eigenfunction of the operator $-\Delta + W^n$, where for brevity we denote by

$$W^n = -\frac{Z}{|x|} + \mu^n \star \frac{1}{|x|} + V^n, \tag{9.65}$$

we may assume (in view of the regularity of W^n, which is in $L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$), that $\phi_1^n > 0$ everywhere. In addition, we use the argument of Section 2, formula (9.24), to claim that, λ_2^n being the second eigenvalue of $-\Delta + W^n$, we have

$$\big((-\Delta + W^n)\theta, \theta\big) - \lambda_1^n \int_{\mathbb{R}^3} \theta^2 \geq (\lambda_2^n - \lambda_1^n) \left(\int_{\mathbb{R}^3} \theta^2 - \big(\int_{\mathbb{R}^3} \theta \phi_1^n \big)^2 \right) \tag{9.66}$$

for any function $\theta \in \mathcal{D}(\mathbb{R}^3)$.

This being done, we pass to the weak limits in the minimizing sequence to obtain some μ, V, ϕ_i, λ_i (as in Step 3) such that

$$(-\Delta + W)\, \phi_i = \lambda_i \phi_i \tag{9.67}$$

where of course $W = -\dfrac{Z}{|x|} + \mu \star \dfrac{1}{|x|} + V$. Moreover, we have $\phi_1 \geq 0$ as a consequence of $\phi_1^n > 0$, and, for any $\theta \in \mathcal{D}(\mathbb{R}^3)$,

$$\big((-\Delta + W)\theta, \theta\big) - \lambda_1 \int_{\mathbb{R}^3} \theta^2 \geq (\lambda_2 - \lambda_1) \left(\int_{\mathbb{R}^3} \theta^2 - \big(\int_{\mathbb{R}^3} \theta \phi_1 \big)^2 \right) \tag{9.68}$$

as a consequence of (9.66). We then rule out the case when one the ϕ_i is identically zero using $J_\varepsilon^{OEP} < I$ and arguing as above. Therefore, we deduce $\phi_1 > 0$ by Harnack inequality on (9.67), and thus ϕ_1 is the ground state of $-\Delta + W$ (as $W \in L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$, the ground state is non-degenerate). We now turn to λ_2 and ϕ_2, which we know is not zero. We know $\lambda_2 \geq \lambda_1$. Suppose $\lambda_2 = \lambda_1$. Then $\phi_2 = \alpha \phi_1$ for some constant α, because of again the nondegeneracy of the ground state. In fact, $\alpha \neq 0$ because $\phi_2 \not\equiv 0$. The fact that $\displaystyle\int_{\mathbb{R}^3} \phi_i \phi_j \leq 1$ in the sense of symmetric matrices writes

$$\left(1 - \int_{\mathbb{R}^3} \phi_1^2\right)\left(1 - \int_{\mathbb{R}^3} \phi_2^2\right) \geq \left(\int_{\mathbb{R}^3} \phi_1 \phi_2\right)^2,$$

in general and thus in the present case

$$1-(1+\alpha^2)\int_{\mathbb{R}^3}\phi_1^2 \geq 0.$$

On the other hand, the Hartree-Fock energy then takes the particular form

$$\begin{aligned} E^{HF}(\phi_1,\phi_2) &= E^{HF}(\phi_1,\alpha\phi_1) \\ &= (1+\alpha^2)\left(\int_{\mathbb{R}^3}|\nabla\phi_1|^2-\frac{Z}{|x|}\phi_1^2\right) \\ &\geq (1+\alpha^2)\left(\int_{\mathbb{R}^3}\phi_1^2\right) I \\ &\geq I. \end{aligned}$$

We reach a contradiction since by weak convergence $E^{HF}(\phi_1,\phi_2) \leq J_\varepsilon^{OEP} < I$. Therefore, we necessarily have the situation when $\lambda_2 > \lambda_1$ and ϕ_2 is a (possibly not normalized) eigenstate associated to λ_2. It then follows using (9.68) and the argument of Section 2 that λ_2 is the second eigenvalue. It also follows that $\displaystyle\int_{\mathbb{R}^3}\phi_1\phi_2 = 0$ and then we enter Step 4 directly. The proof can then be pursued. ♢

9.6 Do the weak formulations $\widetilde{\text{OEP}}$ allow to recover the OEP problems ?

We present in this final section an argument that shows that the minimizer (ϕ_1,ϕ_2) of the weakly formulated problem $\widetilde{J^{OEP}}$ indeed satisfies the "strong constraint" stated in (9.17), in some sense at least. Unfortunately, our proof only applies to the *radial* case (i.e. to problem $\widetilde{J_r^{OEP}}$, together with its complex valued analogue, that we do not detail here for brevity). In addition, we need to assume that the non negative constant c arising in the commutation condition (9.19) and in inequality (9.23) is positive for the minimizer. The additional limitation of our argument is that it cannot be extended to cover the case of the $\widetilde{I^{OEP}}$ problem, for which we can only provide formal arguments.

Let us briefly describe our purpose. A pair (ϕ_1,ϕ_2) that satifies the commutation condition

$$\phi_2\Delta\phi_1-\phi_1\Delta\phi_2 = c\phi_1\phi_2$$

formally satisfies the set of "equations"

$$\left\{\begin{array}{rcl} -\Delta\phi_1+\big(\dfrac{\phi_1\Delta\phi_1+\phi_2\Delta\phi_2+c\phi_2^2}{\phi_1^2+\phi_2^2}\big)\phi_1 &=& 0 \\ -\Delta\phi_2+\big(\dfrac{\phi_1\Delta\phi_1+\phi_2\Delta\phi_2+c\phi_2^2}{\phi_1^2+\phi_2^2}\big)\phi_2 &=& c\phi_2 \end{array}\right. \tag{9.69}$$

and thus formally is a pair of eigenvectors of the same Schrödinger type operator $-\Delta + W$ with $W = \dfrac{\phi_1 \Delta\phi_1 + \phi_2 \Delta\phi_2 + c\phi_2^2}{\phi_1^2 + \phi_2^2}$. The difficulty to give a rigorous sense to this formal statement is twofold. First, we have to prove we may legitimately divide by the density $\rho = \phi_1^2 + \phi_2^2$, and second that the potential W is regular enough for the product $W\phi_1$ and $W\phi_2$ to be given a sense. The two facts are of course closely intertwinned as any information on W gives information on the set of zeros of the ϕ_i.

In the case of the $\widetilde{I^{OEP}}$ problem, we are not able to reach this goal. On the other hand, we can provide rigorous arguments for the $\widetilde{J^{OEP}}$ problem (at least in the radial case) by making use of inequality (9.23) if we assume that $c > 0$.

The sketch of the proof is the following. Let us consider a minimizer (ϕ_1, ϕ_2) of $\widetilde{J_r^{OEP}}$, the existence of which is stated in Theorem 9.4.2. First, we prove

Lemma 9.6.1 *Assume $c > 0$. The functions ϕ_1 and ϕ_2 are continuous (except possibly at the origin) and the set of points $\{x \in \mathbb{R}^3,\ / \phi_1(x) \neq 0\}$ is connex.*

As ϕ_1 is radially symmetric and continuous, the support of ϕ_1 therefore is either the whole space $\mathbb{R}^3$, or a ball, or the complement of a ball, or a hollow ball. It also follows from Lemma 9.6.1 that ϕ_1 does not change its sign, and therefore we may suppose that ϕ_1 is positive. Second, we show that we may always assume, without loss of generality, that ϕ_2 is also supported in $\mathrm{Supp}\,\phi_1$.

Lemma 9.6.2 *Assume that $c > 0$ for at least one solution (ϕ_1, ϕ_2) of $\widetilde{J_r^{OEP}}$). Then, there exists a solution of $\widetilde{J_r^{OEP}}$, still denoted by (ϕ_1, ϕ_2), such that $\phi_1 \geq 0$ and* $\mathrm{Supp}\,\phi_2 \subset \mathrm{Supp}\,\phi_1$.

Finally, we conclude the arguments by showing

Lemma 9.6.3 *Let (ϕ_1, ϕ_2) be a solution $\widetilde{J_r^{OEP}}$ satisfying the properties set in Lemma 9.6.2 (*$\mathrm{Supp}\,\phi_2 \subset \mathrm{Supp}\,\phi_1$*). Let us denote by $\rho(x) = \phi_1(x)^2 + \phi_2(x)^2$ the electronic density and by $\Omega = \{x \in \mathbb{R}^3,\ \rho(x) > 0\}$. Then the potential*

$$W = \begin{cases} \dfrac{\phi_1 \Delta\phi_1 + \phi_2 \Delta\phi_2 + c\phi_2^2}{\rho} & \text{on } \Omega \\ +\infty & \text{elsewhere} \end{cases}$$

is in $H^{-1}(\omega)$ for any open set $\omega \subset\subset \Omega$, and so are the products $W\phi_1$ and $W\phi_2$.

As we shall only work in this section with radially symmetric functions, say ϕ, we shall often make the slight abuse of notations consisting in denoting by $\phi(r)$ the single value $\phi(x)$ for any $x \in \mathbb{R}^3$ such that $|x| = r$.

Proof of Lemma 9.6.1. Since ϕ_i is a radially symmetric function in $H^1(\mathbb{R}^3)$, ϕ_i is continuous except maybe at the origin. Clearly, since $\phi_1 \not\equiv 0$ we may consider some x_0 such that $\phi_1(x_0) \neq 0$, and, say, $\phi_1(x_0) > 0$. Let $r_0 = |x_0|$. By continuity, we may consider the largest $0 < \alpha \leq r_0$ and the largest $0 < \beta \leq +\infty$ such that $\phi_1 > 0$ on $B^c_{r_0-\alpha} \cap B_{r_0+\beta}$. Clearly, if $r_0 - \alpha > 0$ then $\phi_1(r_0 - \alpha) = 0$, and if $\beta < +\infty$ then $\phi_1(r_0 + \beta) = 0$. Suppose that $r_0 - \alpha > 0$ and $r_0 + \beta < +\infty$ (otherwise the following proof is even simpler, since no cut-off is needed at the origin, and/or the cut-off at infinity can be treated likewise). The idea is to consider a sequence of radially symmetric functions $\psi_n \in \mathcal{D}(\mathbb{R}^3)$ such that ψ_n goes to the characteristic function of $B^c_{r_0-\alpha} \cap B_{r_0+\beta}$. This can easily be done for instance by setting $\psi_n(r) = 0$ for any $r \leq r_0 - \alpha$ or $r \geq r_0 + \beta$, $\psi_n(r) = 1$ for any $r_0 - \alpha + \frac{1}{n} \leq r \leq r_0 + \beta - \frac{1}{n}$, $0 \leq \psi_n(r) \leq 1$ for $r_0 - \alpha \leq r \leq r_0 - \alpha + \frac{1}{n}$ or $r_0 + \beta - \frac{1}{n} \leq r \leq r_0 + \beta$, and $\frac{1}{n}\|\psi_n\|_{C^1}$ uniformly bounded with respect to n. Then we pass to the limit in

$$\int_{\mathbb{R}^3} \phi_1^2 \, |\nabla \psi_n|^2 \geq c \left(\int_{\mathbb{R}^3} \psi_n^2 \, \phi_1^2 - \big(\int_{\mathbb{R}^3} \psi_n \phi_1^2 \big)^2 \right), \tag{9.70}$$

in order to obtain

$$0 \geq c \left(\int_{B^c_{r_0-\alpha} \cap B_{r_0+\beta}} \phi_1^2 - \big(\int_{B^c_{r_0-\alpha} \cap B_{r_0+\beta}} \phi_1^2 \big)^2 \right), \tag{9.71}$$

which clearly implies, since $c > 0$ by assumption,

$$\int_{B^c_{r_0-\alpha} \cap B_{r_0+\beta}} \phi_1^2 = 1,$$

and therefore concludes the proof : the total mass of ϕ_1^2 being one, ϕ_1 is therefore supported in the annular (possibly the ball if $r_0 - \alpha = 0$) $[r_0 - \alpha, r_0 + \beta]$.

In order to go from (9.70) to (9.71), we simply have to remark that for any function such as ϕ_1 in $H^1_r(\mathbb{R}^3)$, we have

$$\begin{aligned} |\phi_1(a+t) - \phi_1(a)| &= \left| \int_a^{a+t} \frac{\partial \phi_1}{\partial r} \, dr \right| \\ &\leq \left(\int_a^{a+t} |\nabla \phi_1|^2 \right)^{1/2} \left(\int_a^{a+t} \frac{dr}{r^2} \right)^{1/2} \\ &= \left(\int_a^{a+t} |\nabla \phi_1|^2 \right)^{1/2} \frac{1}{a} \, O(\sqrt{t}), \end{aligned}$$

and therefore, applying this to $a = r_0 - \alpha$,

$$\begin{aligned} \int_{r_0-\alpha}^{r_0-\alpha+\frac{1}{n}} \phi_1^2 \, |\nabla \psi_n|^2 &\leq C \, n^2 \int_{r_0-\alpha}^{r_0-\alpha+\frac{1}{n}} \phi_1^2 \\ &\leq C \, n^2 \left(\int_{r_0-\alpha}^{r_0-\alpha+\frac{1}{n}} |\nabla \phi_1|^2 \right)^{1/2} O(\frac{1}{n^2}) \\ &= o(1), \end{aligned}$$

indeed goes to zero as n goes to infinity. The same applies to the cut-off at $r_0 + \beta$ and the proof of the Lemma is completed. $\diamondsuit$

Proof of Lemma 9.6.2. We now intend to show that we may always assume, without loss of generality, that ϕ_2 is also supported in Supp ϕ_1, namely $[r_0-\alpha, r_0+\beta]$, that we henceforth denote by $[r_1, r_2]$, where we recall that r_1 may be zero, and r_2 may be $+\infty$. We for instance do the proof of the fact that we may always assume $\phi_2(r) = 0$, when $r \geq r_2$ and $r_2 < +\infty$.

To begin with, we make some remarks.

By definition of r_2, we know that $\phi_1(r) \equiv 0$ for $r \geq r_2$. Changing ϕ_2 into $-\phi_2$ if necessary, we may always assume $\phi_2(r_2) \geq 0$. Moreover, we may then change ϕ_2 into the function, still denoted by ϕ_2,

$$\phi_2(r) = \begin{cases} \phi_2(r), & \text{when } r \leq r_2, \\ |\phi_2|(r), & \text{when } r \geq r_2, \end{cases} \tag{9.72}$$

without changing anything in the properties of the pair (ϕ_1, ϕ_2). We henceforth assume $\phi_2 \geq 0$ for $r \geq r_2$.

On the set $r > r_2$, we have the Euler-Lagrange equation

$$-\Delta\phi_2 - \frac{Z}{|x|}\phi_2 + (\phi_1^2 \star \frac{1}{|x|})\phi_2 = -\lambda_2\phi_2. \tag{9.73}$$

Equation (9.73) can be obtained directly on the minimization problem by considering variations of ϕ_2 only on the set $r > r_2$ that keep c fixed (the constraints (9.19) and (9.23) do not play any role for $\phi_1 \equiv 0$ on this open set). It follows from (9.73) that $\Delta\phi_2 \in L^2(B^c_{r_2})$ and that, together from the nonnegativity of ϕ_2 on the same set, we have

$$\phi_2(r) > 0, \text{ when } r > r_2.$$

by Harnack inequality.

Clearly, two cases may occur : $\phi_2(r_2) = 0$ or $\phi_2(r_2) > 0$.

We first show that necessarily $\phi_2(r_2) = 0$. Let us argue by contradiction and assume $\phi_2(r_2) > 0$. Then the strict positivity of ϕ_2, already true for $r > r_2$ can be slightly extended around r_2, by continuity of ϕ_2. More precisely, we may find an interval $[r_2 - \eta, r_2 + \eta]$, $\eta > 0$, where ϕ_2 is bounded below, away from zero by a constant $a > 0$. On such an interval (upon which $\phi_2 \geq a > 0$), we may write the commutation condition as

$$- \div (\phi_2^2 \nabla f) + c\phi_2^2 f = 0$$

with $f = \dfrac{\phi_1}{\phi_2} \geq 0$. We are now allowed to use Harnack inequality to conclude that

$$\sup_{[r_2-\eta/4, r_2]} f \leq \alpha \inf_{[r_2-\eta, r_2]} f$$

for some positive constant α. As we have assumed $\phi_2(r_2) > 0$, $f(r_2) = 0$ and then $f = 0$ on $[r_2 - \eta/4, r_2]$ and we reach a contradiction.

We are now in the situation when $\phi_2(r_2) = 0$, and of course, for $r > r_2$, $\phi_1(r) = 0$ and $\phi_2(r) > 0$. Let us decompose the Hartree-Fock energy of (ϕ_1, ϕ_2) as follows

$$E^{HF}(\phi_1, \phi_2) = A_1 + A_{r_2} + A_{r_2^c} + D, \tag{9.74}$$

with

$$A_1 = \int_{\mathbb{R}^3} |\nabla \phi_1|^2 - \frac{Z}{|x|}\phi_1^2$$

$$A_{r_2} = \int_{B_{r_2}} |\nabla \phi_2|^2 - \frac{Z}{|x|}\phi_2^2$$

$$A_{r_2^c} = \int_{B_{r_2^c}} |\nabla \phi_2|^2 - \frac{Z-1}{|x|}\phi_2^2$$

(note the $\dfrac{(Z-1)}{|x|}$ instead of the $\dfrac{Z}{|x|}$ because the potential generated by ϕ_1 is accounted for), and

$$D = \iint_{B_{r_2}\times B_{r_2}} \frac{\phi_1^2(x)\phi_2^2(y)}{|x-y|} - \iint_{B_{r_2}\times B_{r_2}} \frac{\phi_1(x)\phi_2(x)\phi_1(y)\phi_2(y)}{|x-y|}.$$

Of course, if ϕ_2 is identically zero outside B_{r_2} there is nothing to be proven, so we can assume $\displaystyle\int_{B^c_{r_2}} \phi_2^2 > 0$, and introduce $\mu = \dfrac{\int_{B_{r_2}} \phi_2^2}{\int_{B^c_{r_2}} \phi_2^2}$.

We next consider pairs of the form $(\phi_1, \widetilde{\phi_2} = \alpha\phi_2|_{B_{r_2}} + \beta\phi_2|_{B^c_{r_2}})$ which automatically satisfy the constraints of problem $\widetilde{J_r^{OEP}}$ as soon as we impose

$$\alpha^2 \int_{B_{r_2}} \phi_2^2 + \beta^2 \int_{B^c_{r_2}} \phi_2^2 = 1.$$

We have

$$\begin{aligned} E^{HF}(\phi_1, \widetilde{\phi_2}) &= A_1 + \alpha^2 A_{r_2} + \beta^2 A_{r_2^c} + \alpha^2 D \\ &= E^{HF}(\phi_1, \phi_2) + (\alpha^2 - 1)(A_{r_2} + D - \mu A_{r_2^c}). \end{aligned}$$

We may now choose α first such that $\alpha^2 - 1 > 0$, and next such that $\alpha^2 - 1 < 0$ (note both cases are possible precisely when $\displaystyle\int_{B^c_{r_2}} \phi_2^2 > 0$). This shows that necessarily $A_{r_2} + D - \mu A_{r_2^c} = 0$, otherwise we would contradict the fact that (ϕ_1, ϕ_2) is a minimizer. But then this shows that, for any α and β we have $E^{HF}(\phi_1, \widetilde{\phi_2}) = E^{HF}(\phi_1, \phi_2)$, and

therefore in particular we may choose $\beta = 0$, and leave the Hartree-Fock energy unchanged. Consequently, it is indeed possible to assume that ϕ_2 vanishes outside B_{r_2}.

Proof of Lemma 9.6.3. Let $\omega \subset\subset \Omega$. As $\rho > 0$ on Ω, there exists some positive constant a such that $\rho \geq 0$ on ω. It follows that $f_1 = \dfrac{\phi_1}{\rho}$ belongs to $H^1(\omega)$; indeed, $f_1 \in L^2(\omega)$ and

$$\begin{aligned} |\nabla f_1| &\leq \frac{|\nabla \phi_1|}{a} + \frac{|2\phi_1^2 \nabla \phi_1 + 2\phi_1 \phi_2 \nabla \phi_2|}{\rho^2} \\ &\leq \frac{3}{a}|\nabla \phi_1| + \frac{1}{a}|\nabla \phi_2|. \end{aligned}$$

The same results holds for $f_2 = \dfrac{\phi_2}{\rho}$. Therefore

$$W = f_1 \Delta \phi_1 + f_2 \Delta \phi_2 + c\frac{\phi_2^2}{\rho} \ \in \ H^{-1}(\omega).$$

Moreover, $f_1\phi_1$ and $f_2\phi_1$ also are in $H^1(\omega)$, since a simple calculation shows that

$$|\nabla(f_1\phi_1)| \leq \frac{4}{\sqrt{a}}|\nabla \phi_1| + \frac{1}{\sqrt{a}}|\nabla \phi_2|, \quad |\nabla(f_2\phi_1)| \leq \frac{3}{2\sqrt{a}}\left(|\nabla \phi_1| + |\nabla \phi_2|\right).$$

The product $W\phi_1$ then is well defined in $H^{-1}(\omega)$. Similarly, $W\phi_2 \in H^{-1}(\omega)$.

References

[1] E. Cancès et al., *Computational quantum chemistry : a primer*, in C. Le Bris Ed., **Computational chemistry, a special volume of the Handbook of Numerical Analysis**, Ph. G. Ciarlet, Ed., North-Holland, Amsterdam, to appear 2003.

[2] R.M. Dreizler and E.K.U. Gross, **Density functional theory**, Springer, 1990.

[3] T. Grabo, T. Kreibich, S. Kurth, and E.K.U. Gross, *Orbital functionals in density functional theory : the optimized effective potential method*, in V.I. Anisimov Ed., *Strong Coulomb correlations in electronic structure calculations : beyond the Local Density Approximation*, Gordon and Breach, Amsterdam 2000, 203 - 311.

[4] J.B. Krieger, Yan Li and G.J. Iafrate, *Construction and application of an accurate local spin-polarized Kohn-Sham potential with integer discontinuity : Exchange-only theory*, Phys. Rev. A 45 (1992) 101-126.

[5] P.L. Lions, *Solutions of Hartree-Fock Equations for Coulomb systems*, Comm. Math. Phys. **109** (1987) 33-97.

[6] M. Reed and B. Simon, **Methods of modern mathematical physics. IV. Analysis of operators**, Academic Press 1978.

[7] J.D. Talman and W.F. Shadwick, *Optimized effective atomic central potential*, Phys. Rev. A **14** (1976) 36-40.